2022

Flexible goal-directed manipulation of representations: computational models of healthy and pathological human cognition

Granato, Giovanni

http://hdl.handle.net/10026.1/19674

http://dx.doi.org/10.24382/661
University of Plymouth

UNIVERSITY OF
PLYMOUTH

Flexible goal-directed manipulation of representations: computational models of healthy and pathological human cognition

by

Giovanni Granato

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing, and Mathematics

September 2022

Acknowledgements

Whenever I have the opportunity, I personally declare my gratitude to all the people who contribute to my well-being and, directly or indirectly, to this thesis. Probably almost none of them will read these acknowledgements, so this text may not be useful for these people.

Probably the thesis acknowledgements are intended to support a personal analysis of the PhD period, with the aim of remembering that social support is very important for a researcher. Furthermore, the acknowledgements should highlight that the research is not an 'isolated garden' but a 'shared park' that can only thrive if all the people take care of the park's flora and fauna, together.

Obviously, these considerations can be generalised to many jobs and daily-life situations. I'm autistic and I have personally faced the particular challenge that the social environment can represent for humans, and in particular for autistic people. I have worked hard to enter the social world and, despite being a particular complicated world that often exhibits illogical and harmful dynamics, I think that 'the game is worth the candle'.

For these reasons I have chosen to write these acknowledgements. I will probably forget someone but if he/she knows me and have my respect, they will ignore this mistake with no regrets.

I think it is more efficient to make a list, in no specific order:

- My Psychotherapist Silvana. She guided me trough the social world, giving me the tools to become aware of my condition and my strengths and weaknesses. Moreover, she encouraged me to try to adopt the point of view of other people.

- All animals that I have met over these years. They always give me joy and a chance to poke around in their life

- The dogs of Tiburtina (my city neighbourhood). Every time i met a dog, I experience a magical moment: no social barriers and a lot of shared affection. This habit earned me the nickname 'the dog lover' in my city neighbourhood but I think dogs appreciate my my pampering and attention, while giving me a great dose of joy.

- My parents Brunella and Felice. They have supported me in so many ways in my life and over the years: I think no more words are needed to describe a son's gratitude for parents.

- My girlfriend Mariateresa. She supported me in the complicated stressful situations of a career as a researcher, reminding me that a person's life encompasses many aspects. She also gave me love, intellectual stimulation and respect without which I think I would have had a different path. Overall, both my career, my mind

and my heart benefit and have benefited from her support.

- My supervisors, and in particular Gianluca Baldassarre. They supported me with their trust and suggestions, often on more general aspects of life.

- Many people in LOCEN/LENAI and ISTC. We have shared 7 years of experience and they have influenced my life in many ways over these years.

- All the people I met in many places around the world (eg cafes, bars, train stations). I have often had interesting conversations with strangers and many of them have given me the opportunity to analyse a different point of view.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This study was carried out in collaboration with the Istituto di Scienze e Tecnologie della Cognizione (ISTC) - Consiglio Nazionale delle Ricerche (CNR), Rome.

Relevant scientific conferences were attended at which work was presented and multiple journal articles were published or are under publication.

Publications:

Granato G, Cartoni E, Da Rold F, Mattera A, Baldassarre G (2022) Integrating unsupervised and reinforcement learning in human categorical perception: A computational model. PLoS ONE 17(5): e0267838.
DOI: https://doi.org/10.1371/journal.pone.0267838

Baldassarre, G. and Granato, G. (2022). A Neuro-Computational Theory of Consciousness based on the Internal Manipulation of Representations. Psychological Review. Under revision.
DOI: https://doi.org/10.48550/arXiv.1912.13490

Granato, G., Borghi, A. M., Mattera, A., Baldassarre, G. (2022). A computational model of inner speech supporting flexible goal-directed behaviour in Autism. Scientific reports, 12(1), 1-15.
DOI: https://doi.org/10.1038/s41598-022-18445-9

Granato, G. and Baldassarre, G. (2021). Internal manipulation of perceptual representations in human flexible cognition: A computational model. Neural Networks, 143, 572-594.
DOI: https://doi.org/10.1016/j.neunet.2021.07.013

Granato, G., Borghi, A. M., and Baldassarre, G. (2020). A computational model of language functions in flexible goal-directed behaviour. Scientific reports, 10(1), 1-13.
DOI: https://doi.org/10.1038/s41598-020-78252-y

Baldassarre, G. and Granato, G. (2020). Goal-Directed Manipulation of Internal Representations Is the Core of General-Domain Intelligence. Journal of

Artificial General Intelligence, 11(2), 19-23.
DOI: https://doi.org/10.2478/jagi-2020-0003

Presentations and conferences attended:

Granato G. Baldassarre G. (2022). Manipulation of internal representations underlying flexible human goal-directed behaviour: supporting Computational Psychiatry and towards Machine Consciousness. Poster session presented at "The symposium: from cortical microcircuits to consciousness (CORTICON)".

Granato, G. and Baldassarre, G. (2019). Goal-directed top-down control of perceptual representations: A computational model of the Wisconsin Card Sorting Test. In 2019 Conference on Cognitive Computational Neuroscience (pp. 2019-1168).

Refereeing activity

Reviewer for 'Neural Networks' (2022).

Reviewer for 'Conference on Cognitive Computational Neuroscience' (2019).

Word count for the main body of this thesis: **39880**

Signed: _____

Date: 05/10/2022

Abstract

Title: 'Flexible goal-directed manipulation of representations: computational models of healthy and pathological human cognition'

Candidate: Giovanni Granato

The brain is able to manipulate itself, adapting its internal world in a changing environment. In particular it continuously manipulates its representations to achieve goals. This competence is supported by many neurocognitive high-order processes that interact during the performance of a goal-directed behaviour (e.g. attention processes, executive functions, motivational systems). Overall, adult humans often face a new problem trough a change of the 'perceptual point of view' (i.e. a representational change), rather then an extended research of the correct action to perform. On the other hand, infants and children contemporary develop motor competence and task-directed perceptual representations (e.g. categorical perception). Here I approach the main research question 'how the brain *manipulates* its representations to solve a task that requires cognitive flexibility?'. Moreover, I started to approach the second related research question 'how the brain *acquires* suitable representations to solve a categorisation task?'. First, adopting a synergistic theoretical and computational approach, I identified the systems and basic computational principles that allow the brain to learn, to generate and to manipulate its internal representations in order to achieve a goal. Second, I built a set of computational models and I tested them against experimental human data extracted from already published experimental works. This translational approach corroborates my theoretical proposals and, vice versa, provides scientific and clinical knowledge on the investigated processes. In particular, my models represent a novel computational tool for the investigation of flexible cognition and categorical perception in case of clinical populations (e.g. autistic people). Moreover, they represent a starting point to propose a new theory of conscious cognition showing both scientific implications (e.g new models of consciousness) and technological implications (e.g consciousness-inspired robots).

Contents

List of Tables

List of Figures

Note on the thesis and publications

The theoretical and computational works presented in this thesis have been published at international peer-reviewed conferences and journals.

The candidate is the first author of all publications listed below. Indeed, he has personally carried out all the key steps that have led to these publications such as literature reviews, computational models development, data analysis and interpretations, drawing of figures/plots/schemes, manuscript writing and revision until final publication. Supervisors and other co-authors have supported these projects, providing suggestions at every stage.

Chapter 2 and 3 are adapted from Granato & Baldassarre (2021), Granato et al. (2020), Granato et al. (2022b) and Baldassarre & Granato (2020). In particular, chapter 2 is adapted from the theoretical sections of the published papers (e.g., literature reviews and theoretical proposals) while chapter 3 is adapted from the computational sections of the published papers (e.g. experimental task/conditions, models descriptions, results, discussions). Chapter 4 is adapted from Granato et al. (2022a) and Baldassarre & Granato (2022). In particular, section 4.1 is adapted from the first and section 4.2 is adapted from the second paper. Finally, section 4.2.3 ('Implications of the RIM framework') focuses on similar topics that I have approached in Baldassarre & Granato (2020), i.e. 'representations manipulation in AI'.

Chapter 1

Introduction

1.1 Aim of the work and research approach

Through evolution the brain evolved the capacity for the generation and achievement of goals (Daw et al., 2011; Balleine & Dickinson, 1998). Importantly, the brain manipulates its internal representations in an environment that continuously changes.

Here I highlight that the concept of 'goal-directed manipulation of representations' is approached in isolation by may scientific fields with many different terms such as 'executive functions', 'selective attention', 'internal attention', 'imagination'. I propose that these terms refer to a form of 'self-directed internal action' which aims to modify the agent's conditions, so as to allow it to change the environmental conditions in a goal-directed fashion. This competence is supported by many high-order processes that interact during the performance of a goal-directed behaviour (Diamond, 2013). For example attentional mechanisms extract features from external inputs (selective attention) or activate sensory cortices from an internal input (imagination; Mechelli et al., 2004). Moreover, executive functions and motivational systems participate in key goal-directed processes (Diamond, 2013) such as goal storing (working memory), switching (cognitive flexibility) and monitoring (planning). On the other hand, human perceptual systems ex-

hibit a functional structure that supports the influence of both motivational and attentional processes, hence producing and storing adaptive representations (e.g., visual working memory; Sanada et al., 2013). The importance of goal-directed perceptual manipulation goes further to neuropsychological tasks and impacts on the daily life. Indeed, since adult humans have adequate motor competences, they often solve a new problem trough a change of the 'perceptual point of view' (i.e. a representational change; Duncker, 1935), rather than an extended research of the correct action to perform. On the other hand, infants and children develop motor competence and task-directed perceptual representations (e.g., categorical perception; Goldstone & Hendrickson, 2010; Carvalho & Goldstone, 2016) at the same time, the second of which could constitute an important target for the goal-directed manipulation in adulthood. In this sense, a slow representation learning in childhood and a subsequent fast goal-directed manipulation in adulthood could represent key human abilities that allow us to flexibly overcome the daily challenging problems.

Interestingly, the scientific investigation of human perceptual processes and other cognitive processes supporting goal-directed behaviour (e.g. executive functions) is approached separately by different research fields. This 'knowledge compartmentalisation' is a common approach of neuro-cognitive sciences because the vast amount of details and variations that cognitive processes show (i.e. inter-individual and intra-individual differences). However, this tendency often causes biases toward specific aspects of cognition limiting a more systemic point of view.

For example, computational models of executive functions (e.g., cognitive flexibility; Levine & Prueitt, 1989; Dehaene & Changeux, 1991; Berdia & Metz, 1998; Amos, 2000; Kaplan et al., 2006; Bishara et al., 2010; Caso & Cooper, 2017, 2020; Steinke et al., 2020a,b) consider the key components of cognitive systems and focus on the solution of specific neuropsychological tasks, but often ignore the contribution of the perceptual competences to the task solution. On the other hand, a bulk of psychological studies (Boutonnet & Lupyan, 2015; Foerster et al., 2020; Whorf, 2012; Casasanto, 2008; Dove, 2018; Lupyan & Clark, 2015; Borghi

et al., 2019; Alderson-Day & Fernyhough, 2015) investigate the contribution of language in high order cognitive processes but few computational studies investigate the contribution of inner-speech (a self-directed form of language) to perception and flexible cognition (Garagnani et al., 2008; Garagnani & Pulvermüller, 2013; Cangelosi et al., 2000; Lupyan, 2005; Mirolli & Parisi, 2006; Caligiore et al., 2010). Unfortunately, none of them emulates the contribution of inner-speech to the solution of a cognitive flexibility task or the influence of language as a self-directed manipulation of high-order and perceptual representations. Again, there are many experimental studies (Carvalho & Goldstone, 2016; Witzel & Gegenfurtner, 2016; Wakita, 2004; Maier et al., 2014; Holmes et al., 2009) and computational models (Spratling & Johnson, 2006; Kröger et al., 2007; Salminen et al., 2009; Casey & Sowden, 2012; Pérez-Gay et al., 2017; Tajima et al., 2016; Beer, 2003) that investigate the emergence of category-based perception, but no model attempts to investigate the synergies between motivation and categorical perception, both fundamental for children during the tasks solution.

Considering the main topic of this project and the literature limitations, here I approach the main research question 'how does the brain *manipulate* its representations to solve a task that requires cognitive flexibility?'. Moreover, I start to approach the second related research question 'how does the brain *acquire* suitable representations to solve a categorisation task?'.

First, adopting a synergistic theoretical and computational approach, I identified the systems and basic computational principles that allow the brain to learn, to generate and to manipulate its internal representations in order to achieve a goal. Second, I developed a set of computational models that I tested with a neuropsychological task. In particular, I validated these models with human experimental data that I extracted from already published experimental works. This approach corroborates my theoretical proposals and, vice versa, provides scientific and clinical insights. The following sections detail each phase of this research project.

1.2 Project organisation

This PhD project is composed of three sequential phases, namely 'Main phase', 'Application phase' and 'Post-doc phase'. Figure 1.1 proposes an overview of these phases and their contents.

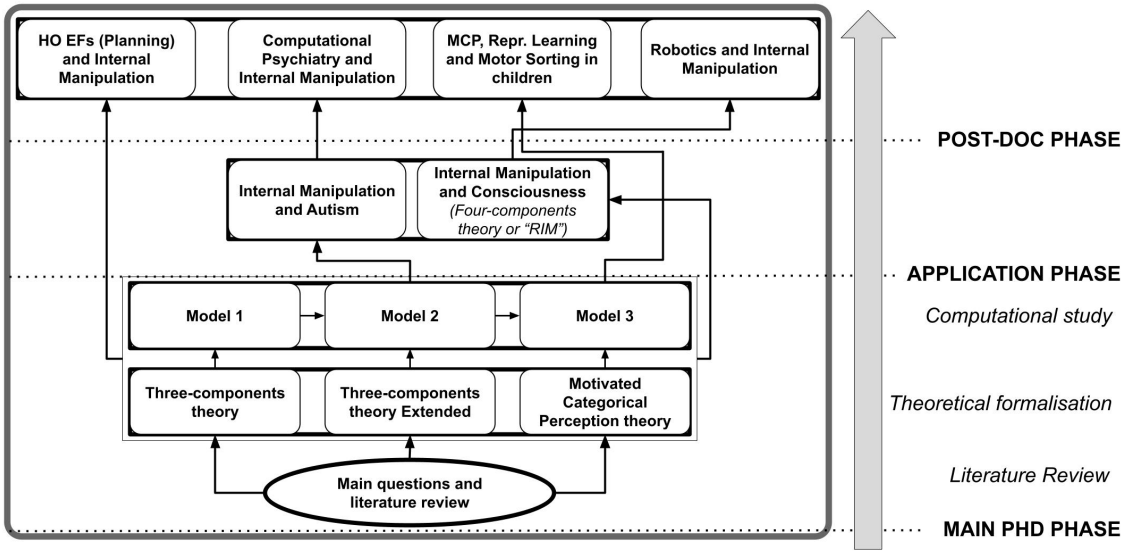


Figure 1.1: Overall schema of the PhD project. HO EFs: High-order Executive Functions. MCP: Motivated Categorical Perception. RIM: Representations Internal Manipulation.

The first phase ('Main') is further divided into three sub-phases, i.e. 'literature review', 'theoretical formalisation', and 'computational study'. The first introduces the main questions of the project and a first focused literature review regarding the project topics (e.g., goal-directed behaviour, executive functions, internal representations, perceptual categorisation, etc). The second proposes formal and computational theories that give a clear description of the investigated processes (i.e. three-components theories and motivated categorical perception theory). The third operationalises and corroborates the proposed theories through the development and validation of computational models (i.e. model 1, model 2 and model 3).

The second phase ('Application') represents a first attempt to go beyond the main studies, proposing a computational application (i.e. inner speech contri-

bution and Autism) and a theoretical deepening of the previous theories (i.e. the four-component theory or 'Representations Internal Manipulation' theory of consciousness).

The third phase ('Post-doc') takes inspiration from the previous two phases to propose many future directions of this project. In particular, it proposes to develop further links between the concepts of human internal manipulation and motivated categorical perception to many fields such as visual planning, computational psychiatry, perceptual/motor skills in children. Furthermore, this phase proposes to investigate the role of internal manipulation as a technological application in robotics.

I now describe in detail each stage of this project.

First, in chapter 2 I analyse the literature regarding the key concepts of my work (e.g., executive functions and categorisation) and in section 2.2 the computational models of flexible cognition and categorical perception. Then, I formalise the role of perceptual processes during a flexible behaviour, proposing specific theories of cognition that are coherent with experimental and clinical studies (section 2.3). In particular, I propose 'the three-components hypothesis', formalising the key brain macro-systems supporting the execution of a flexible behaviour on the basis of the internal manipulation of perceptual representations. This hypothesis describes the processes supporting the performance of a neuropsychological task measuring cognitive flexibility, but both the hypothesis and the task performance can be generalised to daily-life human flexible cognition. In section 2.3 I extend the three-components hypothesis, including the manipulation of high-order representations (e.g. high-order working memory) performed by inner-speech, a self-directed form of language. Together, the three-components theory and its extension corroborate the idea that self-directed representations manipulation is a key factor of flexible behaviour in humans. Moreover, in section 2.4 I propose a 'motivated categorical perception hypothesis' that describes the emergence of categorical perception (CP; i.e. a slow category-based adaptation of perceptual representations that can occur during category learning tasks) as a result of the interaction between per-

ceptual, motivational and motor systems. Note that while the three-component hypothesis describes a "one-shot manipulation of representations", this second theoretical proposal contemplates a slow learning process that allows the acquisition and modulation of adaptive perceptual representations. Interestingly, both the three-components and the motivated CP hypothesis expect motivational systems to drive a manipulation or a slow acquisition of perceptual representations. In this sense, the motivated CP hypothesis could describe the infants and children learning processes preceding the development of a mature self-directed manipulative competence in the adulthood. Future investigations will test the idea that there is a developmental progression from a slow representation learning to a fast self-directed manipulation.

On the basis of my theoretical proposals, I built three neuro-inspired computational models that operationalise them. These models reproduce and explain behavioural data I extracted from already published and validated experimental works, involving different human populations during the performance of neuropsychological tasks (chapter 3). My models development is driven both by the cognitive neuroscience/neuropsychology literature and by recent advances in machine learning (e.g., generative models). Note that even if these models partially share the computational components (e.g., a deep generative model), they show different scopes, partially different architecture, different dynamics, different results, and different future directions. For these reasons this thesis proposes a specific section for each model.

The first model (section 3.1) shows a systemic brain-inspired architecture which is able to store/update specific sub-goals and manipulate its perceptual representations in a goal-directed way. The second model (section 3.2) represents an update version of the first one, showing the addition of an "inner-speech component" that manipulates the stored high-order representations of sub-goals. In this sense, the second model is able to execute a self-directed manipulation of its perception and a second-order manipulation that indirectly influence the first one. The third model (section 3.3) shows a neuro-inspired systemic architecture and it is able to

execute a motor output and, on the basis of its performance, to modulate its motor and perceptual representations. This adaptation is supported by a slow learning process based on the interaction of three components emulating human perceptual, motivational and motor processes. Moreover, emulating representation learning processes in the brain, the perceptual component of this model is trained trough an integration of associative mechanisms and motivational-guided reinforcement learning mechanisms.

Each model performs a task and produces experimental data. I compared the model results with those I extracted from already published and validated experimental works, involving different human populations during the solution of the task. In particular, the first model is tested with a neuropsychological task measuring cognitive flexibility, the Wisconsin Card Sorting Test (WCST; Dehaene & Changeux, 1991; Heaton et al., 2000). Moreover, the results are compared with those obtained from two already published experimental works that involved two healthy populations (young adult and old adults) and two pathological populations (patients with frontal lesions and Parkinson patients). The second model is tested with the same task. In this case the results are compared with those obtained from one already published experimental work that involved young adults populations in different experimental conditions. In particular, the model fits the data obtained during the solution of the same neuropyschological task of the first model with the addition of a verbal shadowing protocol, able to disrupt the advantage that inner-speech can be for participants. At last, the third model is tested with a simple categorisation task. Its results are qualitatively compared with those obtained by many already published experimental works describing the categorisation processes of clinical populations. In particular, different learning profiles of the perceptual component are able to explain the individual differences of the perceptual competence in autism spectrum conditions.

Overall, my models corroborate the theoretical hypothesis for which a flexible switch between different brain representations is a key function for performing a flexible goal-directed behaviour. Moreover, they represent a novel computa-

tional tool for the investigation of flexible cognition and categorical perception in case of clinical populations. For example, in section 4.1 I demonstrate that the second model is able to explain the experimental data extracted from already published papers involving autistic populations. In particular, the model predicts that autistic people show a reduced inner-speech contribution during problem solving. On the other hand, the model predicts that the inner-speech contribution increases during the life span in neurotypical people. In section 4.2 I propose a theoretical deepening of my theories and models, linking them to consciousness and proposing many technological implications. In particular, in section 4.2.3 I investigate the implication of my theoretical and computational proposals in robotics and machine consciousness.

In section 5.2 I present future investigations that aim to overcome the models limitations and to further develop the main research topics. In particular, future directions expect many extensions of my models, the development of new models and the exploitation of my models for clinical and technological scopes.

Chapter 2

Literature review and working hypotheses

In this chapter I introduce a focused literature review with the aim to extract the project key concepts from different scientific fields (e.g. cognitive neuroscience, neuropsychology, computational modelling) and to support my theoretical hypothesis (Figure 2.1). These proposals represent a formalisation of the key investigated processes and working hypothesis for the development of the following computational models.

Although this section reports a bulk of studies from different fields, section 2.1 and section 2.2 provide a general overview of these studies while section 2.3 and section 2.4 relate this literature with my theoretical proposals.

2.1 Goal-directed flexible cognition, internal representations and categorisation

A crucial step in the evolution of brain was the acquisition of goal-directed processes allowing more flexible and complex behaviours with respect to the existing rigid stimulus-response mechanisms. In particular, habitual behaviour ('model-free' behaviour in computational literature; Sutton et al., 1998) is supported by

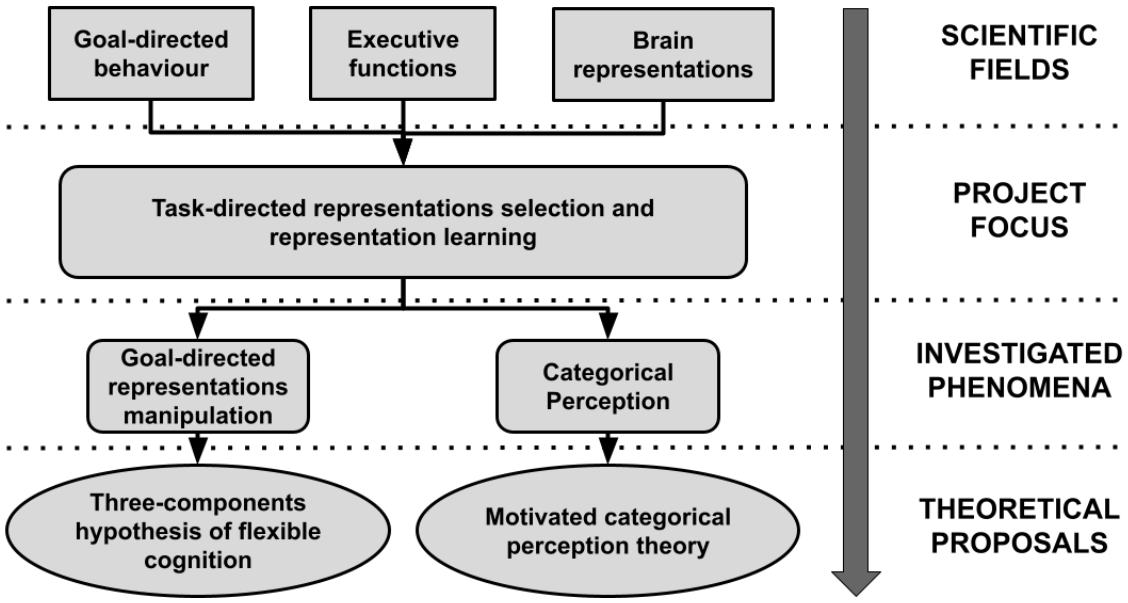


Figure 2.1: Schema showing the workflow of this section, from key topics (e.g. goal-directed behaviour) to theoretical proposals (e.g. three-component hypothesis of flexible cognition).

direct associations between sensations and responses (Gläscher et al., 2010; Yin & Knowlton, 2006a) while goal-directed behaviour (GDB; 'model-based' behaviour in computational literature; Sutton et al., 1998) is able to exploit decision-making processes, flexibly creating on the fly associations between world representations and actions sequence (Daw et al., 2011; Balleine & Dickinson, 1998). The definitions of goal-directed behaviour are partially overlapped to those of executive functions (EFs), a set of top-down goal-directed neuropsychological processes supported by top-down attention. The modern model of executive functions (Diamond, 2013) includes two basic processes, namely the 'working-memory update' - the ability to update the representations stored in working-memory, and the 'inhibition control' - the ability to inhibit interfering responses. Their integration supports cognitive flexibility - the capacity of switching between different behavioural strategies depending on external/internal conditions - and other high-level executive functions (e.g., planning).

Both literatures regarding GDB and EFs investigate the brain processes those lead to the solution of a task, referring to the generation and selection of adaptive representations to reach goals (e.g., the generation of a detailed world representation

or a representations update/switching supporting cognitive flexibility).

Interestingly, there are other research fields that overtly investigate the organisation and coding of brain representations. For example, many studies (Bauer & Just, 2017; Peters et al., 2017; Arbib, 2008) highlight that brain representations can progressively acquire different levels of abstraction , being stored in unimodal perceptual working memories and multimodal abstract working memory (Belger et al., 1998; Quak et al., 2015; D'Esposito, 2007). Other studies particularly focus on the relationship between representations organisation and their functional role underpinning many high–order processes such as categorisation, behavioural switching and planning. For example, Wilcox et al. (2008) theorise that infants, executing circular manipulatory experiences, generate representations based on what they perceive to be relevant. They suggest that this generative process facilitates the visual object segregation and the interpretation of upcoming events. Moreover, Chelazzi et al. (2013) propose that visual selective attention, influenced by reward-based mechanisms, shapes perceptual representations thus providing planning processes with the most efficient representations of the world. Interestingly, Seger & Miller (2010a) suggest that a categorization task requires a balance between reward-shaped slow cortical plasticity and subcortical fast plasticity. This balance should support the generation of a trade-off between generalizable and specific representations. At last, Martin (2016) highlights that embodied cognition requires the generation of distributed representations, integrating both motor, emotional and sensory aspects during the solution of cognitive tasks.

In this project I focus on two examples of task-directed representation learning/modulation: the 'categorical perception' and the a novel concept I defined 'goal-directed representations manipulation'. Now I explain these two phenomena in detail.

'Categorical perception' (CP; Goldstone & Hendrickson, 2010; Carvalho & Goldstone, 2016) is an adaptive learning of sensory representations, possible occurring during a categorisation task, which leads to an increase of the between-category representation differences and a decrease of the within-category representation

difference. Experimental evidence suggest that these learning processes can occur in a bottom-up way, depending on the experienced input patterns (de Zilva & Mitchell, 2012; Wang et al., 2012; Wills et al., 2004), and in a top-down way, depending on task-dependent feedback signals (Caras & Sanes, 2017; Li et al., 2004; Witzel & Gegenfurtner, 2016). However, there is controversial evidence regarding visual stages that show a CP effect. For instance, Wakita (2004) show that CP influences early stages of sensory processing (e.g., V1) while Maier et al. (2014) propose that later cognitive stages of processing support CP (e.g., linguistic labels). Moreover, Holmes et al. (2009) empirically corroborate the idea that both striate and extrastriate cortices support CP. Reconciling controversial results, Ahissar & Hochstein (2004) propose that perceptual learning processes occur at different stages of visual hierarchy, depending on the task demands. In addition, several studies (Lim et al., 2014; Seger & Miller, 2010a) suggest that sub-cortical structures that support reward-based feedback signals (e.g., basal ganglia) interact with cortical structures, contributing to the emergence of category-based perception. In particular, many studies (for an extended review see Lim et al., 2014) suggest that dopamine-based reinforcement learning signals could affect category-related activations in visual sensory cortices.

The concept of 'goal-directed representations manipulation' refers to many forms of one-shot alteration of representations. In particular, I propose this concept to include all processes that, on the basis of a specific goal, change/warp/activate/select specific high-order or perceptual representations. There are several proposals that can be included into this definition.

For example, experimental literature attest that top-down attention processes, supported by frontal/parietal cortices and basal ganglia, perform a modulation on the perceptual representations coming from sensory occipital, temporal, and parietal cortices (Gazzaley & Nobre, 2012; Chelazzi et al., 2013; Zanto et al., 2011). This modulation supports the extraction of external input features and the selection of relevant information (selective attention). Other studies (Stokes et al., 2009; Zacks, 2008; Kosslyn, 1999) focus on the representations generated by imaginary pro-

cesses, suggesting that they are supported by a persistent goal-directed top-down activation executed by frontal-parietal cortices on sensory cortices, in the absence of an external input. At last, modern views (Mechelli et al., 2004; Kornmeier et al., 2009; Dijkstra et al., 2017) suggest that perception and imagination represent the poles of a continuum and depend both on the internal conditions of the agent and the external conditions of the environment. Moreover, it is suggested that language can have an important role into these processes (Boutonnet & Lupyan, 2015; Foerster et al., 2020). In particular, these studies highlight how language influences not only high-level cognition (e.g., reasoning and problem solving) but also perception, for example helping humans to better recognise and categorise objects and entities . Language, both in its overt form of spoken utterances and in its covert form as inner speech, has also been proposed to be conceived as a sort of cognitive tool that can extend my memory and support prediction capabilities (Dove, 2018; Lupyan & Clark, 2015; Borghi et al., 2019). At last, current research highlights different functions that inner speech might have, in particular in relation to cognitive control (Langland-Hassan & Vicente, 2018).

2.2 Computational modelling approaches

There are many computational models of categorical perception and executive functions. The first ones focus on different aspects of CP, such as the interaction between low-level and high-level information at different neuronal sites (Spratling & Johnson, 2006), the systems supporting speech production (Kröger et al., 2007), self-organising mechanisms (Salminen et al., 2009), visual competitive hierarchies (Casey & Sowden, 2012), and the effects of supervised signals (Pérez-Gay et al., 2017). Other models investigate Bayesian inferential mechanisms (Tajima et al., 2016) and embodied evolutionary influences (Beer, 2003). Although these models clarify many aspects of CP, none of them focuses on the computational effects caused by an *interaction* between cortical learning mechanisms (mainly unsupervised learning) and sub-cortical learning mechanisms (mainly reinforcement

learning).

On the other hand, many computational model of executive functions generally focus on planning and the Hanoi tower problem (Stewart & Eliasmith, 2011; Zarr & Brown, 2019; Bieszczad & Kuchar, 2015; Donnarumma et al., 2016) while other models, on which I focus my investigations, investigate the cognitive flexibility and the solution of the Wisconsin Cards Sorting Test (WCST; Levine & Prueitt, 1989; Dehaene & Changeux, 1991; Berdia & Metz, 1998; Amos, 2000; Kaplan et al., 2006; Bishara et al., 2010; Caso & Cooper, 2017, 2020; Steinke et al., 2020a,b). These models clarify specific aspects of executive functions (decision-making, response selection, and feedback-dependent learning) but ignore the representational aspects of cognition highlighted by the literature. At last, there are few models of language functions related to the inner speech and attention (Garagnani et al., 2008; Garagnani & Pulvermüller, 2013; Cangelosi et al., 2000; Lupyan, 2005; Mirolli & Parisi, 2006; Caligiore et al., 2010). However, none of them emulates neither the contribution of inner-speech to the solution of WCST nor the influence of language as a self-directed manipulation of high-order and perceptual representations.

This project proposes three computational studies in which I propose a specific model that I compared with the related computational literature (e.g., models of WCST in section 3.1.5, models of languages as cognitive tool in section 3.2.4, models of categorical perception in section 3.3.5). These comparisons highlight that my models represent the state-of-the-art, at the same time making the differences between my models and other models more clear.

2.3 Representations manipulation: a three-fold hypothesis of flexible cognition

The three-component hypothesis of flexible cognition represents the first theoretical proposal of this project. It states that the internal manipulation of representations relies on the interplay among three fundamental brain systems (Figure 2.2):

(a) a component for storing goals and behavioural rules; (b) a component for manipulating perceptual representations based on the goals and behavioural rules; and (c) a component for extracting perceptual representations and possibly for recalling them based on a bias received from the manipulation component.

The overall top-down modulation of internal representations is a brain mechanism that exerts a bias onto the information flows passing through the cortical pathways (Cisek & Kalaska, 2010; Caligiore et al., 2019a). Several studies (e.g., Kosslyn, 1999; Mechelli et al., 2004; Gazzaley & Nobre, 2012; Fuster & Bressler, 2015; Baldauf & Desimone, 2014; Mannella & Baldassarre, 2015a) have suggested that this mechanism supports the extraction of external input features, favours the top-down biased selection of relevant information (selective attention), and allows the internal persistent or maintenance of information in the absence of an external input (working memory). By integrating a vast number of experimental/theoretical studies (Wolters & Raffone, 2008; Miller & Cohen, 2001; Fuster & Bressler, 2015; Corbetta & Shulman, 2002; Gottlieb, 2007) and computational models (Caligiore et al., 2010; Baldassarre et al., 2013a,b; Mannella & Baldassarre, 2015a; Caligiore et al., 2019a), I consider here how the three components can support flexible cognition when performing WCST.

Note that, although I focus my investigations on a neuropsychological task, the WCST shows ecological validity (Chiu et al., 2018), thus adequately reflecting real-life challenges. Indeed, my theory describes a general framework that formalises the key brain processes underlying human flexible cognition.

The more abstract and amodal *executive working memory* (Hartley & Speer, 2000; Braver & Bongiolatti, 2002) relies on frontostriatal networks. Based on motivational drives, this memory stores the overall goal to pursue (e.g., obtaining positive feedback in the WCST) and the possible sub-goals required to accomplish it (e.g., fulfilling the card matching rules). The goal and sub-goals are encoded as perceptual representations according to the abstraction processes in the perceptual cortical hierarchies.

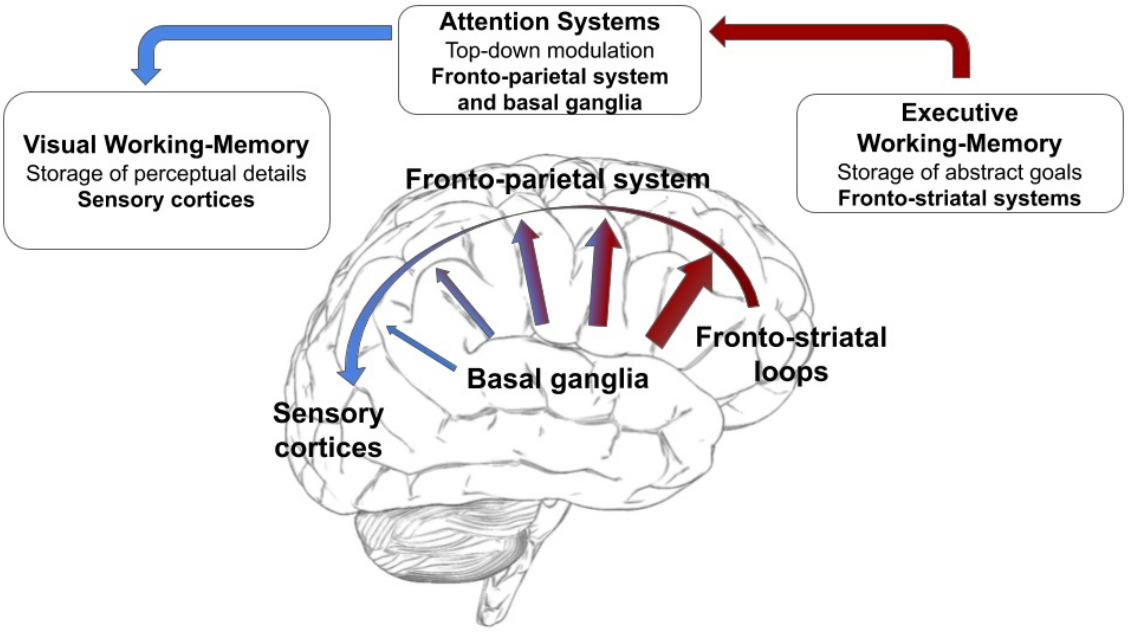


Figure 2.2: Schema showing the three-component hypothesis regarding the internal manipulation of representations. The colour gradient (red to blue) indicates a gradual change in the computational functions from those encoding goals and behavioural rules (red: frontostriatal areas) to those encoding percepts (blue: dorsocaudal areas).

The *frontoparietal cortical system* is based on perceptual attentional processes (Vossel et al., 2014; Parks & Madden, 2013) and basal-ganglia selection mechanisms (Redgrave et al., 1999; Seger, 2008; Chelazzi et al., 2013; Pessoa, 2015). This system applies a top-down bias on the lower-level competition processes that occur within the sensory cortices. In particular, this bias is driven by the behavioural rules stored in the executive working memory and it selects alternative contents within the perceptual cortical system (e.g., the colour rather than the shape of the elements in the card).

The *perceptual cortical system* has two functions. First, it transmits sensory information to the higher-level cognitive systems (Rizzolatti & Matelli, 2003; Gazzaley & Nobre, 2012). Second, it is excited by the top-down biases to implement a perceptual working memory (Raffone et al., 2014) for storing selected perceptual features related to the accomplishment of the required goals (e.g., specific features of cards based on different possible card sorting rules).

Figure 2.3 shows how the three components work in synergy to support the

performance of the WCST. The sub-goals, comprising the behavioural rules linked to specific categories and stored in the executive working memory, stimulate the selection of specific contents in the perceptual working memory. Under this bias, the perceptual working memory generates a representation of the deck card and the target card by emphasising a certain category (e.g., colour). Then the two representations are compared to check whether the two cards match or not. The outcome of this comparison guides the downstream action selection process. Afterwards the system produces the response if the cards match, but if they do not match, another target card is selected to compare with the deck card.

As I highlighted above, these processes could support other tasks and daily-life situations. For example, the same processes could support a visual search task, in which an agent has to search for an object on the basis of an example and/or a specific feature, or a visual plan task, in which an agent has to sort many objects on a table on the basis of an example image.

My hypothesis agrees with previous studies regarding the macrostructure and functional anatomy of the brain. In particular, empirical evidence indicates that the PFC is strongly connected with the parietal cortex to the same extent or even more than with the motor areas (Rizzolatti & Craighero, 2004; Passingham & Wise, 2012). The parietal cortex then exerts strong control on the motor areas. The parietal cortex plays a key role in controlling actions based on representations of the features of objects that are relevant for interacting with them, such as their size and position in space (Jeannerod et al., 1995; Thill et al., 2013). These representations are considered to encode *affordances* (Gibson, 1979; Norman, 1988), that is, the agent's internal representations of the preconditions necessary for the successful accomplishment of actions (Fagg & Arbib, 1998; Thill et al., 2013; Baldassarre et al., 2019a). This idea is at the core of my hypothesis and it contrasts with the view summarised in Figure 2.3 based on all previous models of the WCST. According to these models, the high-level selection of the category rule directly biases the selection of the motor responses rather than the lower-level perceptual representations, as stated in my hypothesis.

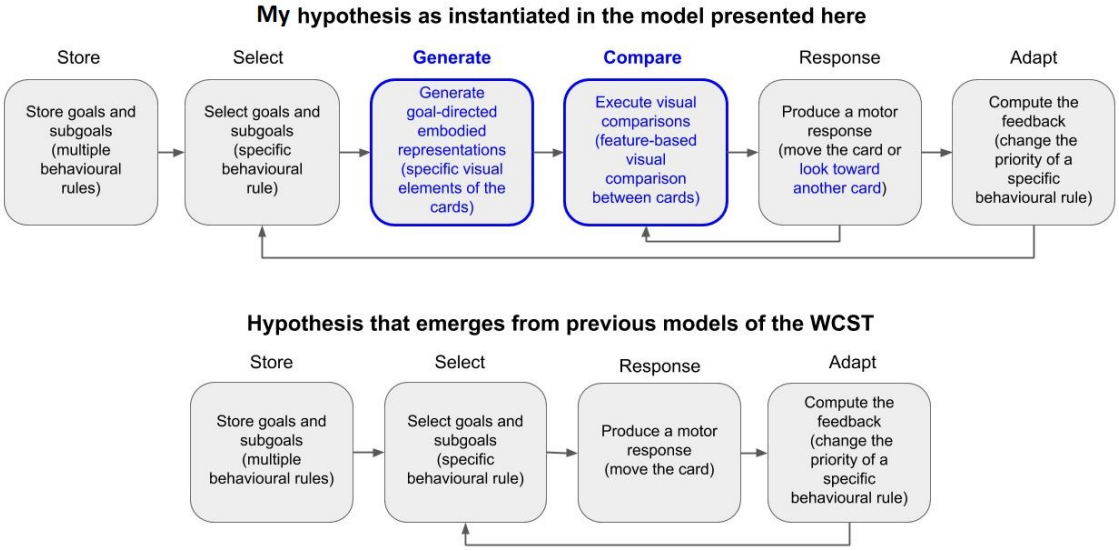


Figure 2.3: Top: Key elements of the proposed hypothesis regarding the processes that might underlie flexible cognition and the solution of the WCST. Bottom: Hypothesis based on other models of the WCST.

A second-order representations manipulation: the case of inner speech

Many brain processes can contribute to shape the internal representations, among them I take in consideration the contribution of a covert form of language contribution, the inner speech.

The notion of inner speech has a long history and is hotly debated in the recent literature (for reviews, see Langland-Hassan & Vicente, 2018; Alderson-Day & Fernyhough, 2015). Inner speech was first introduced by Vygotsky who proposed that it results from a developmental process leading to the progressive internalization of overt speech. Importantly for this work, the notion of inner speech has been later used in the context of working memory, in particular by stressing its role as a rehearsal mechanism that actively maintains information to support planning processes (Baddeley, 1992). Current research focuses on different functions that inner speech might have, in particular in relation to cognitive control (Langland-Hassan & Vicente, 2018) representing a relevant issue also investigated here. Moreover, recent literature attested the existence of different kinds of inner speech, such as wilful/deliberative (Perrone-Bertolotti et al., 2014) vs. spontaneous inner speech, condensed vs. expanded inner speech, monologic vs. dialogic inner speech, and

evaluative/motivational inner speech (Alderson-Day & Fernyhough, 2015). Importantly for this work, some authors have highlighted the relationship of inner speech with second order cognition and metacognition (Clark, 1998) and recent studies have investigated the relation between metacognition and the strategy changes adopted following error detection (Yeung & Summerfield, 2012).

The neural underpinning of inner speech have not been much investigated but many studies proposed specific coherent explanations. For example Geva et al. (2011) suggest that supramarginal gyrus and frontal inferior gyrus, composing the dorsal route of language, are involved into the production of inner speech. In particular, they suggest that the Broca area, and close regions, product inner speech and trasmit it, trough the arcuate fasciculus, to the posterior regions involved in language comprehensions. Similar regions are detected by Hurlburt et al. (2016) that however found an inverse activation pattern between frontal activations, more linked to a deliberate/elicited inner speech, and temporal auditory cortices, more linked to spontaneous inner speech. Loevenbruck et al. (2019) corroborate the involvement of same networks, attesting the activation of left inferior gyrus (Broca's area) and premotor cortices, involved into speech production, and posterior temporal regions, involved in speech comprehension. Lœvenbruck et al. (2018) proposed a sensory-motor account of inner speech. In particular, they propose that parietal and temporal regions transmit a signal to frontal cortices, that produce an inner speech related motor command and so a retro-activation to sensory cortices. At last Marvel & Desmond (2012) suggest that inner speech, supporting the verbal working memory, involves both frontal and parietal cortices. Importantly, they attest that the inner speech-related structures are more activated in case of a requested manipulation of verbal information rather then the simple storing of its, highlighting the role of inner speech in information manipulation.

On the basis of these evidence and the three-component hypothesis, I propose the key idea that the inner speech executes a 'second order manipulation' of internal representations. In particular, the interaction between inner speech structures and dorsal prefrontal cortices (Figure 2.4, on the left) makes the high-order representa-

tions (e.g., goals and sub-goals) more adapt to guide a flexible manipulation of perceptual representations and external behaviour (Figure 2.4, on the right). This view suggest the existence of two goal-directed embodied-manipulation loops (Figure 2.4).

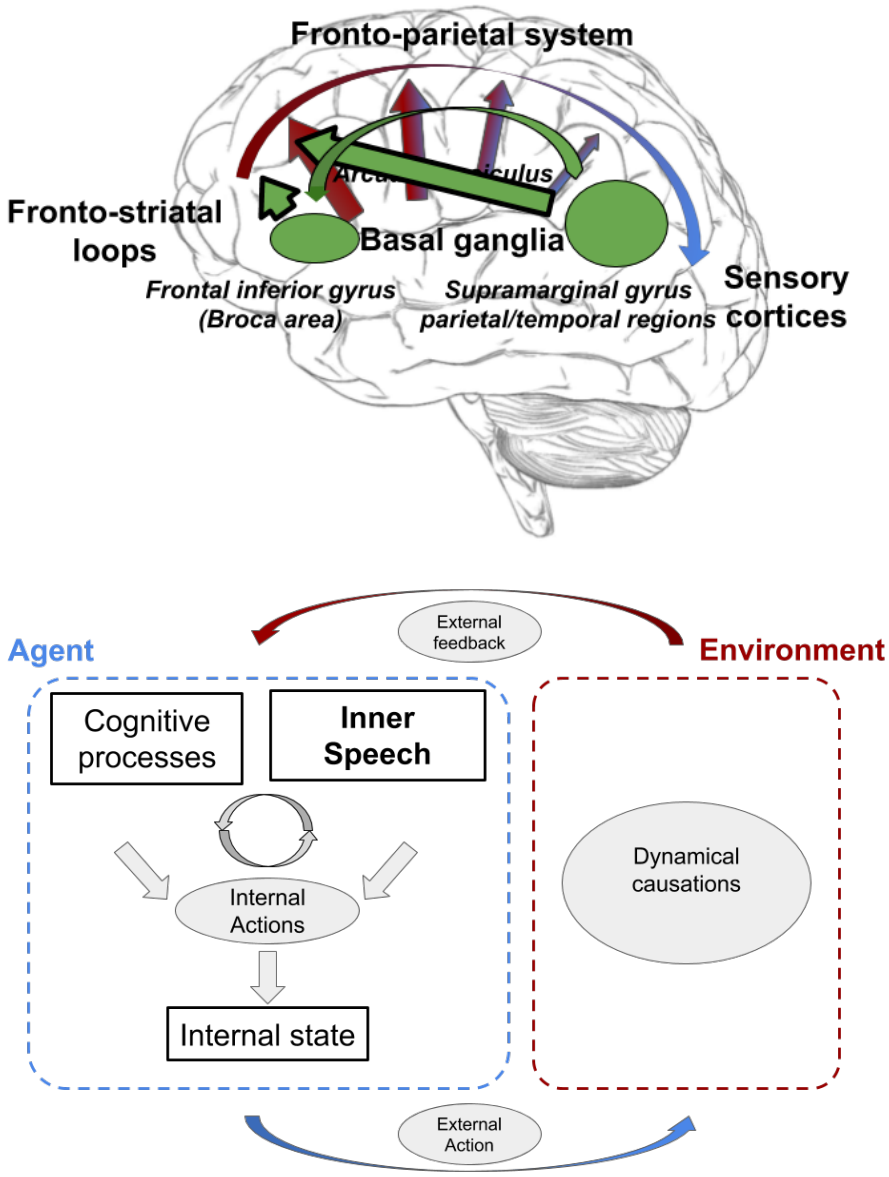


Figure 2.4: On the left: neural correlates of three-fold hypothesis and inner speech production and comprehension. On the right: abstract schema that represents the model double loop of manipulation of internal and external states. The internal manipulation of states, to which inner speech contributes, allows the agent to better manipulate the external environment.

The first loop involves the classic embodied interaction of the agent with external objects (e.g., the visual search and the card manipulation processes performed to solve the WCST) and the first-order manipulation of perceptual representa-

tions (defined in the schema with the term "internal actions"). The second loop involves inner speech in its role of a second-order manipulator, influencing the high-order internal representations (i.e. sub-goals stored in working memory) and so improving the effectiveness of the first loop.

Note the in this context the terms 'first-order representations and manipulations' refer to the perceptual representations and their manipulations, while the terms 'high-order representations' and 'second-order manipulations' refer to the manipulation of abstract and amodal representations of goals and sub-goals stored by executive working memory (dlPFC). Indeed, both terms refer to a self-directed form of manipulation at different levels of abstraction.

2.4 Representations learning processes that precedes a top-down manipulation: the motivated categorical perception theory

The *motivated categorical perception theory* (MCP theory) represents the third theoretical proposal of this project. It formalises the early motivational and sensory-motor processes leading to the emergence of adaptive perceptual representations. In particular, the functional architecture and learning processes proposed by the MCP theory are based on the interactions between brain cortical and sub-cortical macro-systems (e.g. striate/exstrastriate cortices, basal ganglia, motor cortices), supporting specific computational functions (e.g. perceptual abstraction, motivational bias, motor selection).

This theory integrates theoretical proposals and experimental evidence from different research fields that investigate the emergence of brain representations.

For example, fMRI experiments on humans and monkeys show that most cortical regions are activated by reward signals with a trial-locked timing (Pleger et al., 2008, 2009; Vickery et al., 2011; Arsenault et al., 2013) and the dopamine probably

mediates these reward-related signals (Pleger et al., 2009; Arsenault et al., 2013). Furthermore, evidence from discrimination tasks (Pleger et al., 2008, 2009) suggests that the reward-induced reactivation of sensory cortex tunes the representations in a task dependent way. Integrating these evidence, the Superlearning theory (Caligiore et al., 2019b) proposes that different learning processes can coexist in the same brain structure (e.g., associative hebbian and reward-based mechanisms). At last, the theory is also coherent with the theoretical framework of *embodied perception*. Indeed, many studies (Ferdinando & Parisi, 2004; Vernon, 2008; Foglia & Wilson, 2013; Da Rold, 2018) propose that the brain constructs internal representations of the world 'for being ready to act', also establishing a relation between embodied cognition and categorical perception (Collins & Olson, 2014; Schendan & Ganis, 2012; Davis & MacNeilage, 2000). Similarly to the three components hypothesis of flexible cognition, the MCP theory expects that the world influences the perceptual states of the agent, in turn affecting its response. In this case categorical perception emerges in a feedback-dependent manner trough a slow representation learning processes affected by the action of an agent and the clues provided by the world. According to some views (Da Rold, 2018), these elements of perception-action loops between the agent and the environment represent a key feature of embodiment.

On the basis of these experimental evidences and studies on categorical perception, the MCP theory formalises the integration of two features that are shown in Figure 2.5A. First, sensory-motor hierarchy intermediate layers (extra-striate cortices) host mixed UL and RL processes while early layers (striate cortex) and later layers (motor cortices) respectively host unsupervised and reinforcement learning mechanisms. Second, task-dependent signals from the world (i.e. rewards) direct reach the perceptual component. This proposal represents a simplified description of brain processes but it captures the macro differences in learning processes leading to categorical perception. Note that here I specifically focus on the categorical perception and its possible relationship with the existence of UL/RL interactions suggested by the Superlearning hypothesis (Caligiore et al., 2019b). However, my

proposal is only one possible interpretation of a specific group of brain adaptive learning dynamics. Indeed, alternative views propose an higher segregation of the learning modalities in the brain (Doya, 1999) and other modelling approaches emulate the emergence of adaptive dynamics adopting a pure UL approach (Ha & Schmidhuber, 2018) or pure RL approach (e.g. meta-RL; Wang et al., 2018).

Figure 2.5B summarises the main 'agent-environment interactions' that the theory proposes to be at the basis of a categorisation task: perception of the input (bottom-up spread of input information from the world), behavioural response (production of an output toward the world), feedback computation (computation of the external world feedback, e.g. reward signal), and learning (reward-based adaptation of sensory-motor processes).

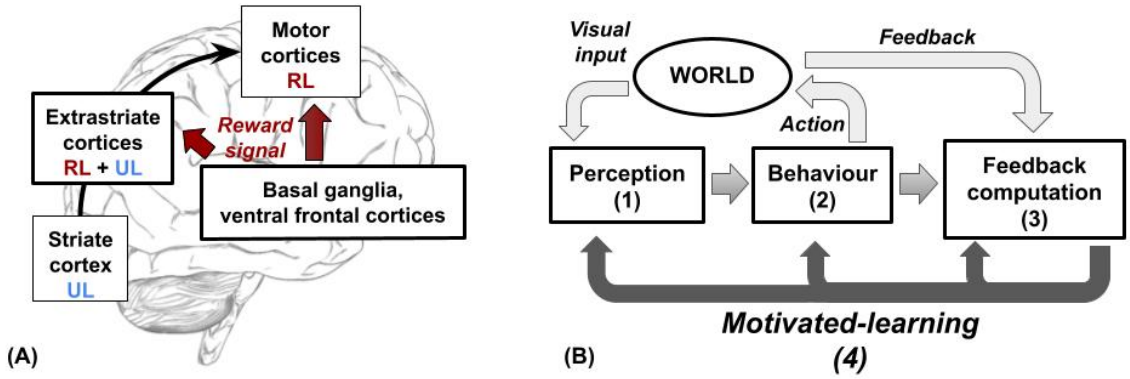


Figure 2.5: (A) Scheme of learning processes and targeted brain areas formalised by the motivated categorical perception hypothesis. The intermediate sensory-motor layers (extra-striate cortices) undergoes both associative learning (UL) and trial-and-error learning (RL). The latter presents a gradient having a decreasing strength moving from the motor cortex towards the striate cortex. (B) A schema of the main model processes involved in its interaction with the environment during the task performance.

The top of the figure highlights a 'sensory-motor loop', in which the agent iteratively perceives the world and executes an action. The bottom of the figure highlights a 'learning loop', for which the agent adapts its sensory computation, behaviour and feedback computation trough a learning process.

Chapter 3

Main computational models and related results

In this section I present three main studies showing the computational models I used to corroborate my theoretical proposals and to explain human experimental data. Each study shows the model architecture and processes, the performed task, the results, a discussion of results and other models. At last each study proposes a brief conclusions section. Note the the conclusions section of each computational study is more focused with respect to the "general conclusions" of the whole project (section 5), which gives an overall view of the project insights.

3.1 Model 1. Top-down manipulation of internal representations to support flexible cognition

Here I introduce the first computational study that corroborates the three-component hypothesis of flexible cognition, focusing on the manipulation of internal representations. In particular, this section introduces the neuropsychological task, the computational components of model and the obtained results. At last I propose a discussion and conclusions about this study.

3.1.1 The Wisconsin Card Sorting Test (WCST) and the brain processes involved in its solution

In the WCST (Figure 3.1, top), the participants must match a card drawn from a deck (called a 'response card' or 'deck card') with one of four sample cards (called 'stimulus cards' or 'target cards'). Each card contains coloured items with a unique combination of features. These features are differentiated into three categories and each has four attributes: (1) colour: red, green, blue, or yellow; (2) form: stars, triangles, circles, or crosses; and (3) number: one, two, three, or four elements. The participant is requested to move the deck card close to one of the target cards by trying to match them in terms of either their colour, form, or number. A first key challenge in the test is that the participant is not told the correct rule for matching the cards. After each action, an operator provides 'correct' or 'incorrect' feedback based on the current matching rule. The participant must then infer the correct rule based on this feedback. A second key challenge in the test involves probing cognitive flexibility. The correct matching rule changes after a certain number of uninterrupted correct actions and when this occurs, the participant must search and switch to the new rule based only on the information provided in the feedback. Finally, in order to pass the test, the participant must complete a certain number of uninterrupted card sequences, where each involves a different rule.

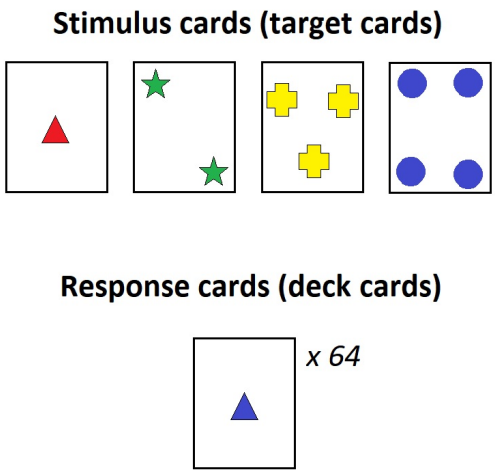


Figure 3.1: Schema showing the typical elements in the Wisconsin card sorting test.

Many versions of the WCST are available with differences in the test procedure or performance score. I used Heaton's version of the test Heaton et al. (2000) with the following specific features: (a) participants can use up to two decks of 64 cards; (b) completing a category set requires ten correct matches in sequence; (c) after completing a category set, the sorting rule changes but the participant is not told; (d) in order to pass the test, the participant must complete a series involving six different correct matching rules: colour, form, number, colour, form, and number; and (e) if the participant uses both decks without completing the series of categories, the test is considered 'failed'.

Scoring and types of errors

To score the test, I followed the official documentation for the test (Heaton et al., 2000). In particular, I used the following five principal indices to give a full profile of the participant's performance (see the documentation for thorough explanations of the indices).

- *Completed Categories* (CC): this index ranges from $(0,6)$ and indicates the number of successfully completed categories to score the global performance.
- *Total Errors* (TE): total incorrect responses, including both perseverative errors and non-perseverative errors (see below), as an index for scoring the level of global deficit.
- *Perseverative Errors* (PEs): cards sorted with the same incorrect rule after a negative feedback error as an index representing perseverative behaviour.
- *Non-Perseverative Errors* (NPEs): errors not included in PEs, where these errors can occur in different situations and they may suggest attentional failure or incorrect inferential reasoning.
- *Failure-to-Maintain Set errors* (FMS): any error that occurs after five consecutive correct matches.

Neural correlates of the behaviour exhibited during the WCST solution

Previous studies have proposed various partially overlapping interpretations of the neural correlates of the behaviour exhibited by WCST participants. Figure 3.2 presents a schematic overview of the brain areas that have a high activation during the performance of the WCST and of lesioned sites linked to specific errors that occur during the test. To build this diagram, I analysed the studies described in the meta-review by Nyhus & Barceló (2009).

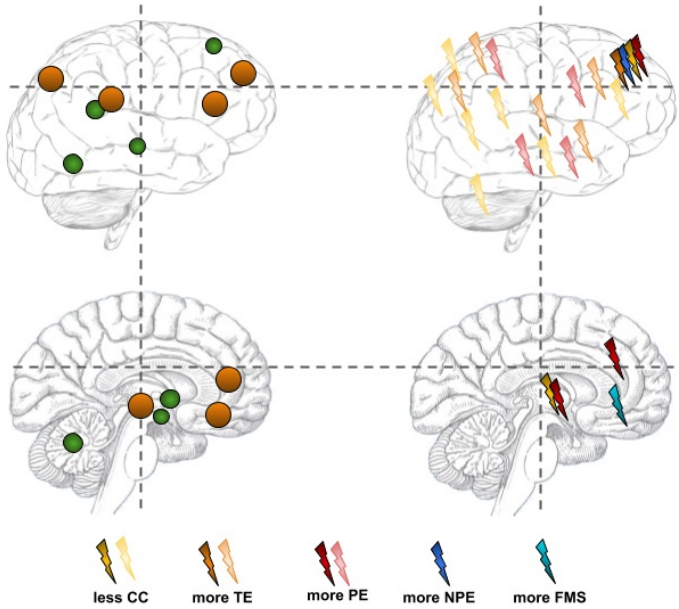


Figure 3.2: Left: Highly active brain areas during the performance of the WCST. The colour and size of each circle indicates the number of studies considered that identified a specific activation site (small/green: < 3; large/orange $\geqslant 3$). Right: Sites of lesions that cause specific errors during the WCST. The colour intensities of the bold arrows indicate the specificity of lesions (transparent: distributed lesions; dark: focused lesions).

This analysis indicates that the brain areas that contribute most to the performance of the WCST are the frontoparietal cortices associated with goal-directed perception and attention (Vossel et al., 2014; Parks & Madden, 2013), sub-cortical structures (particularly basal ganglia) linked with the processing of rewards (Yin et al., 2008), frontal structures such as the orbital and ventromedial prefrontal cortex (PFC) that support emotional processing, and the anterior cingulate cortex (ACC) associated with error detection (Stuss et al., 2000; Zald & Andreotti, 2010). Several studies (e.g., Goldman-Rakic, 1996; Hoffmann, 2013) indicate that these

frontal, parietal, and subcortical systems form an integrated network of systems that underlie cognitive flexibility and other executive cognitive functions.

Figure 3.2 also summarises the relationships between errors that occur during the performance of the WCST and different lesion sites. In agreement with the functional interpretations of the active areas observed in healthy participants discussed above, neuropsychological studies have highlighted the presence of (a) a correlation between PEs and a lesion in both the subcortical and medial cortices, and (b) a correlation between most types of errors and a lesion in the superior frontal cortices. In particular, PEs are considered to be indicators of an impairment in flexibility related to the incapacity to change a behavioural rule that has been successful up to a certain point (Dehaene & Changeux, 1991; Nelson, 1976). In addition to these classical interpretations focused on PEs, some studies focused on NPEs and FMS errors. In particular, Li (2004) suggested that NPEs are related to attentional or reasoning failures, whereas Barceló & Knight (2002) linked them to attention and working memory dysfunctions. Figueroa & Youmans (2013) focused on FMS errors and suggested that they reflect distractibility rather than a cognitive flexibility deficit.

3.1.2 Neuro-inspired underpinnings of the model: key components and dynamics

The architecture of the model was designed based on the organisation of the macro-structural areas of the brain that underlie the functions relevant to my hypothesis: perceptual and category learning, working memory, and the internal selection of representations. The model was abstracted over the anatomical and physiological details of the brain micro-circuits and neurons. This simplification allowed me to realise the first operationalisation of the three-component hypothesis. More biologically plausible implementations of the components might be realised in future work. The architecture also encompasses some auxiliary components that are required for an agent to autonomously form realistic representations of the

cards and to interact with the environment. The architecture and components of the model are shown in Figure 3.3, and they are now explained in detail.

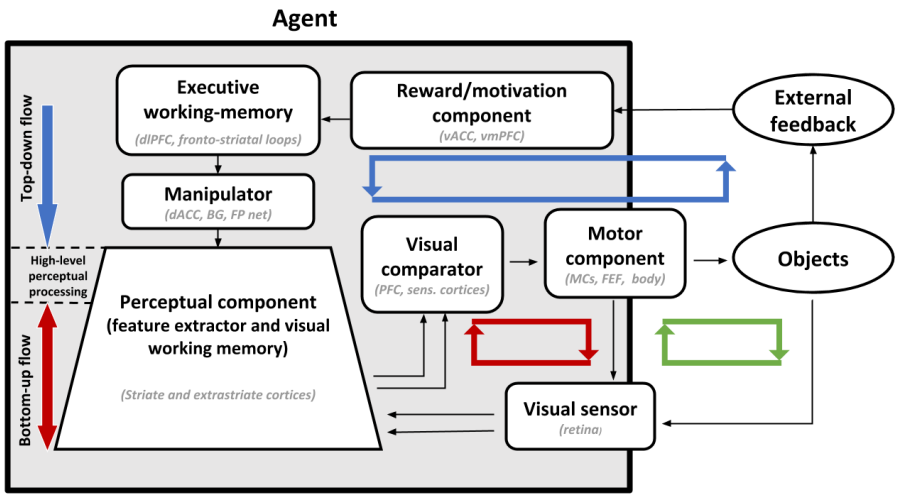


Figure 3.3: Schema showing the model components, functions, flows of information between the components, and interaction loops that allow the agent to engage with the environment (red: attentional loop; green: object-displacement loop; blue: feedback-manipulation loop).

Visual sensor This component corresponds to the retina of the eye. The agent actively displaces the sensor so that it focuses on one card at a time (see below). The sensor returns a visual image of the cards. The image is sufficiently large such that a focused card is completely within its scope.

Perceptual component This component is a layered neural network that performs the bottom-up processing of visual images. The component reflects the hierarchical nature of visual cortices involving many levels of information processing (Felleman & Van Essen, 1991; Mechelli et al., 2004; Baldassarre et al., 2013a) ranging from the low-level retinotopic visual processing of features in the striate cortex (V1) to the processing of higher-level image proprieties (shape, colour, etc.) in extra-striate cortices (V2 to V5) and different areas of the inferotemporal and parietal cortex (DeYoe et al., 1996; Rizzolatti & Matelli, 2003; Konen & Kastner, 2008).

Reward/motivation component This component processes the input to compute the system's internal reward signals and applies a motivational bias to the working memory processes. In the brain, these processes rely on the ventral basal ganglia (Humphries & Prescott, 2010; Mannella et al., 2013), ventromedial PFC, and ventral portion of the ACC (Gläscher et al., 2008, 2012).

Executive working memory This component reproduces the functions of the executive working memory, particularly storing the possible sub-goals that correspond to the possible card matching rules. The executive working memory is supported by dorsolateral portions of the PFC, which can store information regarding goals and behavioural strategies through recurrent circuits (Hartley & Speer, 2000; Braver & Bongiolatti, 2002; Barraclough et al., 2004), and select them based on lateral inhibitory mechanisms (Aron, 2007). Similar to the working memory of the brain (Brunel & Wang, 2001; Gruber et al., 2006), the model executive working memory implements the following key processes (cf. Frank et al., 2001; O'Reilly & Frank, 2006): (a) the active maintenance of sub-goals in the absence of perceiving the corresponding stimuli; and (b) releasing (forgetting) this information when it is no longer relevant.

Perceptual manipulator This component supports two processes. The first process involves decisions regarding behavioural rules that depend on activation of the executive working memory. This process mimics the role of the dorsal ACC and other PFC areas in affective decision making (Bush et al., 2002; Heilbronner & Hayden, 2016; Silvetti et al., 2018). The second process involves the performance of the actual top-down manipulation of the perceptual contents of the perceptual system. The manipulation is based on a disinhibition mechanism that reproduces the main features of the functioning of basal ganglia (Redgrave et al., 1999; Mannella & Baldassarre, 2015a). This mechanism inhibits all internal representations activated by the bottom-up sensory information, except for the one that corresponds to the card matching rule that needs to be followed. Based on this mechanism, only the colour, form, or size features are used to compare the deck and target cards.

The manipulator also employs a local selection process to enhance the activation of specific features observed in the stimuli and in agreement with the top-down bias (e.g., to select 'red' if the chosen behavioural rule is 'colour'). This process mimics the top-down modulation effect of the high-level cortices on the lower sensory cortices via the frontoparietal cortical system and basal ganglia (Gazzaley & Nobre, 2012; Vossel et al., 2014; Parks & Madden, 2013; Yin & Knowlton, 2006a).

Visual comparator This component compares the deck card and the foveated target card, and returns their level of similarity ('visual matching'). This component is inspired by findings related to same/not-same tasks (Perani et al., 1999), which have been shown to rely on the interplay between the occipital/temporal cortices and dorsolateral PFC.

Motor components The first motor component moves the visual sensor to scan the deck card and then the different target cards in order to search for the one that matches the deck card based on the selected matching rule. This component guides the gaze in a top-down manner based on the visual comparator output, as follows. When there is not a visual match, the mechanism triggers a saccade that shifts the fovea to the following target card. When there is a match, the mechanism stops the gaze on the current target card and releases the arm action. When this occurs, the second motor component controls a simulated manipulator that moves the deck card close to the selected target card, as required by the WCST. It is assumed that these attentional scanning and object-moving behaviours are acquired before the solution of the WCST.

Bottom-up perceptual processes and top-down attentional processes

The model implements bottom-up and top-down information flows that correspond to perception and attention processes, respectively (Intaitė et al., 2013; Dijkstra et al., 2017). Perception involves the bottom-up transmission and progressive abstraction of visual information from the retina to higher cortical levels.

Attention and imagination involve a top-down information flow through the frontal areas of the brain that can bias peripheral perceptual areas, and thus they tend to exhibit stronger activation corresponding to relevant external stimuli (attention; Mechelli et al., 2004), or they can even be activated in the absence of them (imagination; Kosslyn, 1999). In the model, the selection of a specific matching rule within the working memory and the consequent disinhibition of a certain attribute representation start a top-down activation flow. This flow leads to the generation of images at the lower levels of the perceptual component that correspond to the selected attribute. In order to perform the task, the model uses these generative processes both with the deck card and the target card under focus. The resulting rule-based representations are then used to compare the two cards at the low perceptual level with respect to the selected colour/shape/size category.

A specific consideration must be made regarding the latter process, as follows. The deck/target card comparison could be performed based on the high-level representations of cards, such as in the last layer of the perceptual component. However, as discussed in Section 2.3, the comparison at the low level is proposed to mimic the functioning of the brain. Moreover, this approach might also have the following computational advantages: (a) the possibility of exploiting the detailed information received from the sensors and selected in a suitable manner by the top-down processes to conduct operations that cannot be performed at a higher level of abstraction (e.g., operations that depend on the detailed shapes of objects; Mechelli et al., 2004; Wolters & Raffone, 2008); and (b) the possibility of using high-level abstract representations to generate lower-level detailed representations based on information gathered 'along the way' while the activation process spreads through the intermediate representation levels, where this more detailed information can then be used by processes that depend on it, such as fine-level comparisons (Gazzaley et al., 2008; Barceló et al., 2000; Mangun, 1995; Woldorff et al., 1997).

Interaction loops

The top-down manipulation mechanism described above is coupled with three interaction loops via the environment (Figure 3.3). These loops allow the model to perceive visual stimuli (images) with a substantial level of realism and, most importantly, to use the manipulated representations to support flexible behaviour.

A first 'attentional loop' involves the motion of the visual sensor, which allows the model to observe deck and target cards, thereby affecting the model's internal processing. A second 'object-displacement loop' involves the motor system, which allows the model to displace the deck card, thereby affecting the following visual percepts. A third 'feedback-manipulation loop' allows the model to process feedback to produce an internal reward signal, which is used to update the relevance of the used matching rule stored in the executive working memory. Thus, the third loop affects the operation of the other two loops.

These loops involving circular interactions between the model components and environment capture the essence of the sensorimotor interactions involved in the solution of the WCST. The loops are simple but sufficient to study the proposed top-down perceptual manipulation mechanism. The model does not have a full embodiment, for example it lacks specific actuators with realistic physical dynamics, but it still has some key embodied features. In particular, the actions of the model can affect its sensory input and the active control of this input is part of the strategy used by the model to perform the task. According to some views with which I agree, the fact that the solution to a problem relies on the circular loop where the agent interacts with the environment represents a key element of embodiment (Nolfi & Floreano, 2000).

3.1.3 Computational details of the model

The key computational features of the model are summarised in Figure 3.4. Algorithm 1 gives an overview of the information flows exchanged by the model

components, the computations executed by these components, and the interactions between the model and the environment.

Algorithm 1 Model: information flows, computations, and interaction loops with the environment.

1: **for** deckCard $\in \{1,2,...,64\}$ **do**
2: (deckCardImage, deckCardPosition) $\leftarrow$ VisualComponentScan(deckCard)
3: attributePreactivation $\leftarrow$ DBNForwardSpreading(deckCardImage)
4: category $\leftarrow$ SoftMax(workingMemoryState)
5: attribute $\leftarrow$ DisinhibitionOfCategoryAttributes(attributePreactivation, category)
6: reconstructedDeckCard $\leftarrow$ DBNGeneration(attribute)
7: match $\leftarrow$ $False$
8: **for** targetCard $\in \{1,2,3,4\}$ AND match = $False$ **do**
9: (targetCardImage, targedCardPosition) $\leftarrow$ VisualComponentScan(targetCard)
10: attributePreactivation $\leftarrow$ DBNForwardSpreading(targetCardImage)
11: attribute $\leftarrow$ DihinibitionOfCategoryAttributes(attributePreactivation, category)
12: reconstructedTargCard $\leftarrow$ DBNGeneration(attribute)
13: match $\leftarrow$ VisualComparision(reconstructedTargCard, reconstructedDeckCard)
14: MotorComponentMoveDeckCard(targetCardPosition, deckCardPosition)
15: feedback $\leftarrow$ GetFeedback()
16: workingMemoryState $\leftarrow$ UpdateWorkingMemory(feedback)

The algorithm involves a first cycle (line 1) where each step corresponds to a card drawn from the deck. In each step of the cycle, the model first visually scans the deck card (line 2). Next, it processes the card features that correspond to the matching rule stored in the working memory and memorises these features for later use (lines 3–6; a non-neural memory is used for this purpose). A second nested loop allows the model to visually scan one target card after the other in each step (line 8). For each target card, the model reconstructs its features corresponding to the current selected matching rule (lines 9–12) and then compares them with those of the deck card stored in memory (line 13). When a target card matches the deck card, the model stops scanning the target cards and moves the deck card below the last scanned target card (line 14). The model then collects the resulting feedback (line 15). Finally, the model uses the feedback to update the working memory (lines 16).

Environment The agent acts in a simulated environment comprising a square space of 100×100 pixels. The environment contains 'objects' (the cards) that the model can visually explore and move in space (see Figure 3.4). The objects are cards representing polygons characterised by a unique combination of three visual properties (categories), where each has one of four possible attributes: colour (red, green, blue, or yellow), form (square, circle, triangle, or bar), and size (large, medium large, medium small, or small). This set of attributes generates $4^3 = 64$ combinations (cards). With respect to the original task, I substituted the 'number' category with the 'size' category because perceiving a different number of objects required higher resolution and this slowed the simulations. For the same reason, I also substituted the form attributes 'stars' and 'crosses' with the attributes 'squares' and 'bars', respectively.

Visual sensor The visual sensor returns a 28×28 pixel RGB image covering a limited portion of the environment. The resulting $28 \times 28 \times 3$ matrix is stored in a vector of 2352 elements that represents the input for the perceptual component. The visual sensor is first directed towards the deck card and then towards the target cards in sequence until the model finds a target card that matches the deck card.

Perceptual component This component is implemented as a deep generative model, specifically a *Deep Belief Network* (DBN; Hinton et al., 2006; Le Roux & Bengio, 2008) comprising two stacked *Restricted Boltzmann Machines* (RBMs; Hinton, 2012). In the following, I explain the main features of the component. An RBM is formed by two layers of units comprising a 'visible' layer and a 'hidden' layer, which are fully connected. A distinctive feature of RBM networks, and thus of the DBN, is that information can flow in both a bottom-up and top-down manner within it. The bottom-up flow of the network (from the visible layer to the hidden layer) reduces the dimensionality of the input pattern (Hinton & Salakhutdinov, 2006) and the top-down flow (from the hidden layer to the visible layer) produces a visible input. The capacity of the network to utilise the activation of the last

hidden layer through a top-down information flow to produce the possible input that corresponds to this activation is an important property called *generativity* (Goodfellow et al., 2017). The perceptual component is trained offline with a novel algorithm, which allows it to extract the specific attributes of each card and represent them in a distributed manner, and thus the model can use the generativity to simulate top-down attention processes. For example (Figure 3.4), I consider a case where the model perceives a 'large, red, triangle' deck-card and a 'medium large, red, square' target card, and the category selected at the higher levels is 'colour'. In this case, the model can lead to the activation of a red blob at the lower levels for each card in the sequence and decide that the two cards match. The bidirectional activation of the component can be repeated many times to simulate the activity reverberations in the visual working memory, thereby allowing me to study the possible loss of information in the presence of interfering distractors if stimulus–response delays are introduced. Given that the participants can freely observe the deck and target cards as many times as they like in the WCST, I assume that there is no loss of information while performing the visual matching of cards. As a consequence, I fixed the number of reverberations of the perceptual working memory to 1 cycle involving a single spreading bottom-up activation followed by a single top-down reconstruction.

Executive working memory The perceptual component is formed by three units encoding the three matching rules (colour, form, and size). The activation of each unit encodes the likelihood that the corresponding behavioural rule is selected. In particular, the units have a continuous value ranging from 0 (low chance) to 1 (high chance), and they store, based on a recurrent self-connection, a representation of the possible matching rules to use. Activation is fuelled by the feedback signal with a binary value from $\{0, 1\}$. The feedback signal only affects the activation of the unit encoding the last selected rule, as follows:

$$m_{s,t} = (1 - \mu) \cdot m_{s,t-1} + \mu \cdot r, \tag{3.1}$$

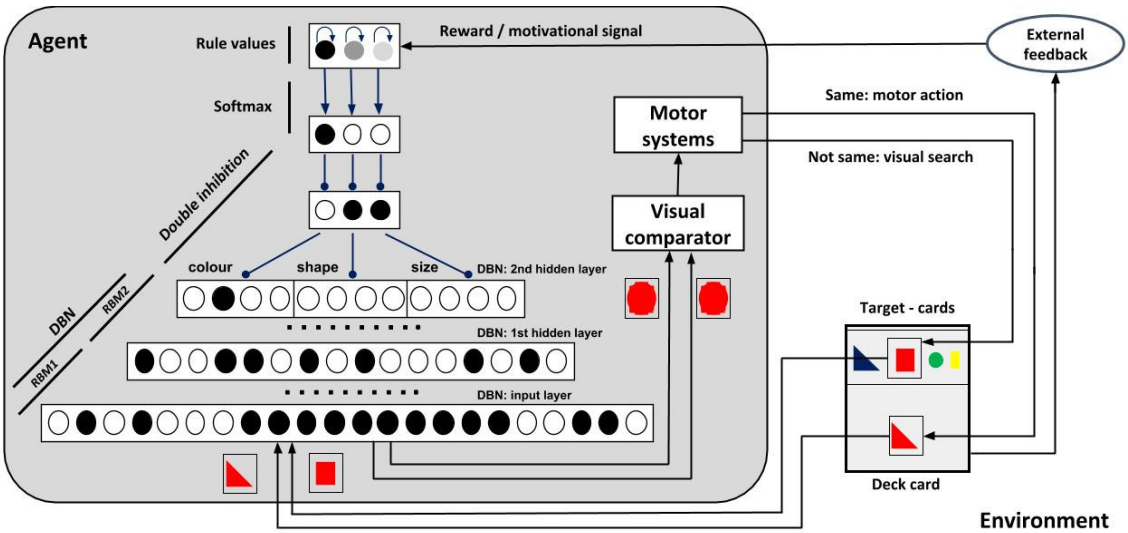


Figure 3.4: Architecture of the model showing the deep belief network for perception, disinhibition mechanism for rule selection, and the rule values and softmax function for matching-rule selection. A stimulus used in the WCST is shown at the bottom right, where the small square frames around the red triangle and the red square represent two 100×100 pixel images corresponding respectively to a deck card and a target card collected by the system visual sensor in successive steps. The two analogous squared frames around the two red circles under the 'visual comparator' are the images obtained by considering the fact that the high levels of the model focus on the 'colour' category and the 'red' attribute to compare the two input cards.

where $m_{s,t}$ is the new activation for the rule unit, $s \in \{1,2,3\}$ is the index for the selected rule, $m_{s,t-1}$ is the previous activation of the unit, $(1-\mu)$ is the strength of the unit recurrent connection, μ regulates the impact of the feedback on the memory, and r is the feedback signal, which is equal to 1 in the case of positive feedback (matching the deck and target cards) and 0 otherwise. In the case of positive feedback, the parameter μ assumes a fixed value of 0.7, whereas in the case of negative feedback, μ is considered to be a free parameter that possibly has different values (see Section 4.1.1 for details regarding the search for the model parameters). I used this approach because previous studies suggest that disengagement (switching after negative feedback) is a critical feature for detecting individual differences and also for pathological behaviours assessed with the WCST (Monchi et al., 2004; Zanolie et al., 2008). The parameter μ is the first of the three key parameters in the model investigated in the simulations.

All non-winning units of working memory decay exponentially towards a baseline

value as follows:

$$m_{l,t} = (1 - \phi) \cdot m_{l,t-1} + \phi \cdot \alpha, \tag{3.2}$$

where $m_{l,t}$ is the value related to the losing unit l ($l \in \{1,2,3\}$; $l \neq s$) at time t, $1 - \phi$ is the strength of the recurrent connection, and α (set to 0.5) is the baseline value to which the memory unit activation converges. A high value of ϕ causes a high rate of information forgetting. The parameter ϕ is the second of the three key model parameters investigated in the present study.

Perceptual manipulator This component implements the following three processes. The first process is a winner-take-all (WTA) competition that receives the values from the working memory as inputs and chooses the matching rule based on the $\mathtt{softmax}$ function:

$$\Pr(k = s) = \frac{\exp(m_k/\tau)}{\sum_{q=1}^{3} \exp(m_q/\tau)}, \tag{3.3}$$

where $\Pr(k = s)$ is the probability of the event that the matching rule k ($k \in \{1,2,3\}$) is selected ($k = s$) and τ is the 'temperature parameter' in the $\mathtt{softmax}$ function for regulating the randomness of the selection. A high value of τ leads to high randomness/exploration of the behavioural rules. The parameter τ is the third of the three key model parameters investigated in the present study. The probabilities $\Pr(\cdot)$ sum up to 1 and they are used to stochastically select the matching rule for use. It should be noted that the stochasticity of the softmax function is the unique source of the behavioural variability of the model. The second process leads the winning unit in the WTA competition to apply a double inhibition mechanism to disinhibit the units of the last DBN hidden layer corresponding to the chosen category. The third process is a localistic winner-take-all mechanism (Srivastava et al., 2013) involving the units encoding the attributes of each category group, and it is applied to the last DBN hidden layer before disinhibition. In this process, the unit with the maximum sigmoid activation (e.g., encoding 'red') is assigned an activation value of 1, whereas the other units (e.g., encoding blue, green, and

yellow) are assigned an activation value of 0. For example, this process can allow the activation of the 'red' attribute if the 'colour' category is disinhibited.

Visual comparator This component computes the Euclidean distance between the two reconstructed images of the deck card and the focused target card. Using a fixed threshold β ($\beta = 0.1$), the component returns a Boolean value representing the result of the comparison ('match'/'not match'). This process is an abstraction of neural comparison processes (e.g., see Santucci et al., 2016).

Motor component This component encompasses two mechanisms. The first mechanism receives the positions of the deck and target cards, and locates the visual sensor (saccades) on them in a sequential manner. This approach captures the essence of more sophisticated attentional mechanisms that the model could use in future studies, such as a bottom-up attention mechanism based on image salient areas or the inhibition-of-return mechanism (Klein, 2000) that I used in previous models (e.g., Baldassarre et al., 2019a).

The second mechanism in the motor component receives the positions of the deck card and the matched target card, and performs a movement to bring the deck card close to the matched target card. This action is hardcoded in my method and it is assumed to be learned by the model before the test (see Baldassarre et al., 2019a).

3.1.4 Results

The results are presented in the following three sections. In Section 3.1.4, I present a validation of the model by showing how it can reproduce the behaviour of healthy and pathological humans in the WCST. In Section 3.1.4, I explore the relationships between the model's behaviour and its key parameters. In particular, I present the results obtained by correlation analysis to investigate the links between the model's parameters and the behavioural indices exhibited by humans in the WCST. In this section, I also present the results of a 'lesion' experiment where the

key model parameters were altered by setting them to extreme values to further investigate the links between these values and the behavioural results. Finally, in Section 3.1.4, I analyse the internal functioning of the three key elements of the mechanism used by the model for the internal manipulation of representations, thereby highlighting its key role in the production of flexible behaviour. The behaviour, underlying reasoning processes, and internal representations of the model can be observed in action in a video at: *https://youtu.be/pnBWWqhULsE*

Validation of the model with human data

I targeted four groups of participants to validate my model (Heaton et al., 2000; Paolo et al., 1995). All of the participants completed the standard version of the WCST (Heaton et al., 2000). In particular, I considered two pathological groups of 59 frontal patients with a local or diffused frontal lesion (average age of 42 ± 14.32 years), and one group of 181 Parkinson patients (average age of 68.92 ± 8.28 years). The pathological group and Parkinson group were paired by age, education, and IQ with control groups of 362 young adults and 162 old adults, respectively.

Model configurations that obtained the best fits to data from healthy and pathological humans I searched for the values of the three key parameters in the model (μ: sensitivity to negative feedback errors; ϕ: working-memory forgetting speed; and τ: exploration/distractibility) using a grid search algorithm (Van Geit et al., 2008). The large number of models tested with this technique also allowed me to use a *sensitivity analysis* (Hamby, 1994) to assess the performance of the obtained parameters. Table 3.1 shows the values of the model key parameters obtained with the automatic search method, that is, the model parameters that resulted in the lowest minimum squared error between the behavioural indices for the human groups and the model groups. Figure 3.5 presents a general view of the model's parameter configurations that obtained the best fits to the four human populations. The plot shows the differences between young healthy participants and the other three populations, thereby supporting the idea that the effect of

ageing on healthy old participants can mimic a frontal impairment (Dennis & Cabeza, 2012; Sullivan et al., 2001).

Among the models used to fit the data reported by Heaton et al. (2000), the 'pathological model' with the parameters that obtained the best fit to the data related to frontal patients had a lower μ, higher ϕ, and similar τ compared with the 'healthy model' that obtained the best fit to the control group (healthy young participants).

Among the models used to fit the data reported by Paolo et al. (1995), the model configurations fitted to the Parkinson patients had lower sensitivity to negative feedback (μ) and higher distractibility (τ) compared with the paired control group. Surprisingly, the Parkinson model had a low working memory forgetting value (ϕ) compared with the related control group (healthy old participants), although it was still higher than that for the healthy younger participants in the study by Heaton et al. (2000). Moreover, healthy old participants had a similar forgetting speed to frontal patients.

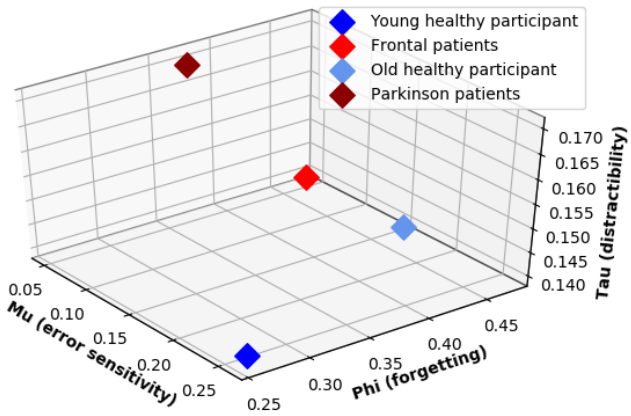


Figure 3.5: Three-dimensional representations of the parameter configurations in the models that obtained the best fits to the four human populations.

WCST indices for the healthy and pathological models, and corresponding human groups I further validated the model versions with the parameter configurations discussed in the previous section by studying the accuracy with which they reproduced the multiple behavioural indexes exhibited by the corresponding

	Error sensitivity (μ)	Forgetting speed (ϕ)	Distract-ibility (τ)
Healthy young participants	0.26	0.26	0.14
Frontal patients	0.05	0.47	0.14
Healthy old participants	0.16	0.47	0.14
Parkinson patients	0.05	0.37	0.17

Table 3.1: Values of the parameters in the models that obtained the best fits to the target WCST data related to the behavioural indices for healthy participants and frontal patients (data from Heaton et al., 2000), and for Parkinson controls and patients (data from Paolo et al., 1995).

groups of human participants. For each model parameter set (human group), I ran and compared 59 simulated participants (that varied by using different seeds in the random number generator) with the 59 frontal participants, and 59 other simulated participants with the 362 healthy participants considered by Heaton et al. (2000). Figure 3.6 shows the average values of the indices for the healthy humans and those obtained by the model, and Figure 3.7 shows an analogous comparison for frontal patients. The comparisons showed that the model reproduced the values of the WCST indexes for all of the human groups considered with high accuracy. Table 3.2 presents the p-values obtained from statistical comparisons of the behavioural indices produced by the models and those for the human groups. The indices were not statistically different ($p > 0.05$), except CC was higher in the two models compared with the human participants (healthy human versus healthy model: 5.18 ± 1.52 vs 5.9 ± 0.4, $p < .01$; pathological human versus pathological model: 3.46 ± 2.25 vs 4.6 ± 1.0, $p < .01$), and FMS was higher in the healthy model compared with the humans (1.4 ± 1.3 versus 0.67 ± 1.09, $p < .01$). These differences were due to the very low variability of the behaviour of the model (see the standard deviations in the figures) leading to a statistically disproportionate weighting on small mean differences. The lower variability of the data obtained by the model could have been caused by the simplicity of the architecture of the model compared with the human brain. In particular, the softmax function (Equation 3.3) is the unique source of variability in the model, whereas the human brain exhibits high variability in terms of its architecture and functioning (participants pursue multiple goals in parallel even when performing

the task, e.g., they might aim to save energy or be socially compliant), thereby leading to individual differences in cognition and behaviour (Finn et al., 2015; Chen et al., 2015; Barch et al., 2013; Kanai & Rees, 2011; Hearne et al., 2016).

In the experiments based on Parkinson patients and the paired healthy control group reported by Paolo et al. (1995), most of the WCST indices obtained by the model were not statistically different from those for the human groups. Again, statistical differences were found only for CC and FMS, where the values were higher using both models because of the same reasons explained above for frontal patients. The statistically non-significant t-tests did not indicate that the results were the same but they further corroborated the capacity of the model to reproduce multiple behaviours of the target human groups.

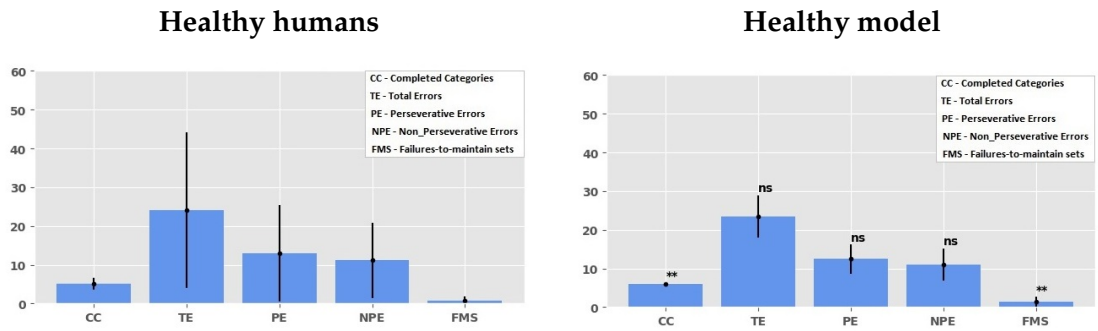


Figure 3.6: Healthy condition: comparison between the healthy model group and healthy human group (** indicates a statistically significant difference at $p < 0.01$).

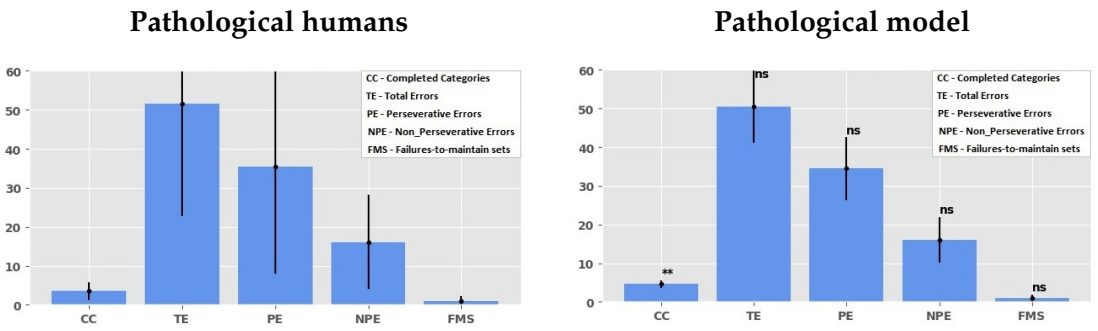


Figure 3.7: Pathological condition: comparison between the artificial impaired group and human frontal patients (** indicates a statistically significant difference at $p < 0.01$).

Participants	Indices				
	CC	TE	PE	NPE	FMS
Healthy (Heaton et al., 2000)	*.001*	.800	.748	.920	*.001*
Frontal (Heaton et al., 2000)	*.001*	.784	.794	.953	.565
Healthy (Paolo et al., 1995)	*.004*	.873	.763	.969	*.003*
Parkinson (Paolo et al., 1995)	*.004*	.678	.635	.964	*.000*

Table 3.2: Statistical comparisons (p-values, two-tailed t-tests) of human data vs. model data involving the healthy and pathological conditions (data from Heaton et al., 2000), and healthy and Parkinson conditions (data from Paolo et al., 1995). The statistically significant p values ($p < 0.05$) are highlighted in *Italics*.

Study of the internal functioning of the model

Relationships between the key parameters of the model and WCST behavioural indices I analysed the relationships between the three key model parameters and the WCST indices by considering their correlations measured using Pearson's *r* (Table 3.3). The analysis showed that CC, indicating the global performance of the model, tended to correlate with high error sensitivity (μ; $r = +0.25$), low forgetting (ϕ; $r = -0.14$), and low distractibility (τ; $r = -0.67$). The analysis also indicated that TE had a negative relation with error sensitivity (μ; $r = -0.60$) and a positive relation with distractibility (τ; $r = +0.34$). The analysis also showed that PE had a robust negative relation with negative feedback processing (μ; $r = -0.58$) but negligible correlations with the other two parameters. The analysis also indicated that NPE had a remarkably positive correlation with distractibility and erratic behaviours (τ; $r = +0.75$), a moderate positive correlation with forgetting (ϕ; $r = +0.24$), and a negative correlation with the error sensitivity (μ; $r = -0.26$). Finally, the FMS errors had a strong correlation with distractibility (τ; $r = +0.73$) and negligible correlations with the other two parameters.

Effects of focused alterations of the model on WCST behavioural indices The simulated lesion technique allowed me to further investigate the role of each single key model parameter in the production of the flexible behaviour measured using the WCST indices. In particular, Table 3.4 shows the three sets of parameters used to obtain the three alternative versions of the control model investigated in this study. The first model called the 'extreme perseverative model' (EPM) is

Indices	Parameters		
	—	ffi	fi
CC	0.25	- 0.14	*- 0.67*
TE	*- 0.60*	0.17	*0.34*
PE	*- 0.58*	0.07	- 0.05
NPE	- 0.26	0.24	*0.75*
FMS	0.04	0.00	*0.73*

Table 3.3: Pearson's r values indicating the correlations between the key parameters in the model (μ, ϕ, and τ) and the different WCST indices. Except for the correlation related to ϕ-FMS, all of the correlations were statistically significant ($p < 0.001$). Correlations stronger than |0.3| are highlighted in *Italics*.

characterised by a very low value for μ. The second model called the 'distracted model' (DM) is characterised by a very high value for τ. The third model called the 'irrational model' (IM) is characterised by a high value for ϕ.

	μ	ϕ	τ
Control model	0.26	0.26	0.14
Extreme perseverative model	*0.001*	0.26	0.14
Distracted model	0.26	0.26	*0.4*
Irrational model	0.26	*1*	0.14

Table 3.4: Parameter values used in the impaired models for producing focused alterations. Values in *italics* represent the altered parameters with respect to the values found by fitting the data of the healthy participants in the study by Heaton et al. (2000).

A global view of the proportions of error with the three altered models (Figure 3.8) confirmed that the EPM and DM had opposite PE/NPE imbalances, where the former had high PEs and low NPEs, and the latter had low PEs and high NPEs. Moreover, the EPM had few FMS errors whereas the DM had many. The IM had slightly more errors than the healthy model, with an imbalance towards NPEs.

Analysis of the functioning of the model mechanism for the internal manipulation of representations

In this section, I describe my investigation of the relationships between the behaviour of the model and the computations of the core components of the three-component hypothesis instantiated in the model, that is, the executive working memory, top-down representation manipulator, and visual working memory. First,

Error profiles for models with lesions compared with the healthy model.

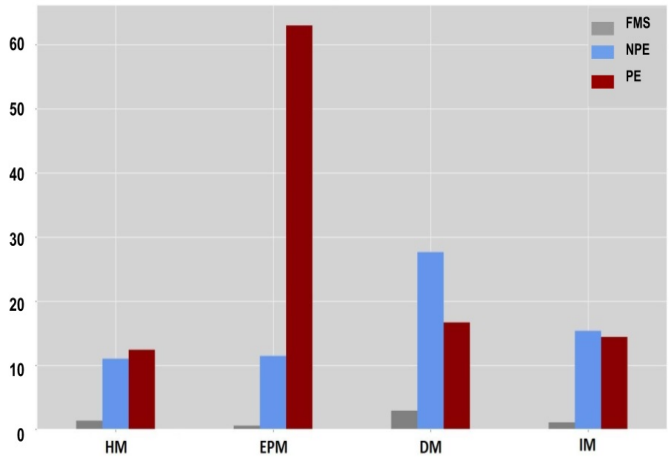


Figure 3.8: Proportion of errors in the altered models compared with the healthy model. HM: healthy model; EPM: extreme perseverative model; DM: distracted model; IM; irrational model; PE: perseverative errors; NPE: non-perseverative errors; FMS: failure-to-maintain set errors.

I studied the relationships between the activation of the executive working memory units and the resulting model actions and errors in healthy and pathological conditions. Next, I studied the internal functioning of the perceptual component, particularly to show how the top-down manipulator based on the sub-goals of the model affected its internal representations of the input stimuli.

Executive working memory dynamics To illustrate the functioning of the working memory component, I plotted the working memory activations and related behavioural responses for the five models. In particular, I considered the healthy model and pathological model with the parameters obtained by fitting the data reported by Heaton et al. (2000) (Figure 3.9), and the three altered versions of the model considered in the previous section, that is, EPM, DM, and IM (Figure 3.10).

In the graphs, the parts of the curves increasing from 0.5 to 1.0 during ten correct responses indicate a successfully completed card category. Curves decreasing from 1 to 0.5 indicate that the desirability of a matching rule decreases, thereby possibly leading to the selection of a different rule. Low values in the two curves followed by the increase in the third curve indicate inferential reasoning by exclusion. Sequences of several small peaks above the baseline (0.5) after a category change

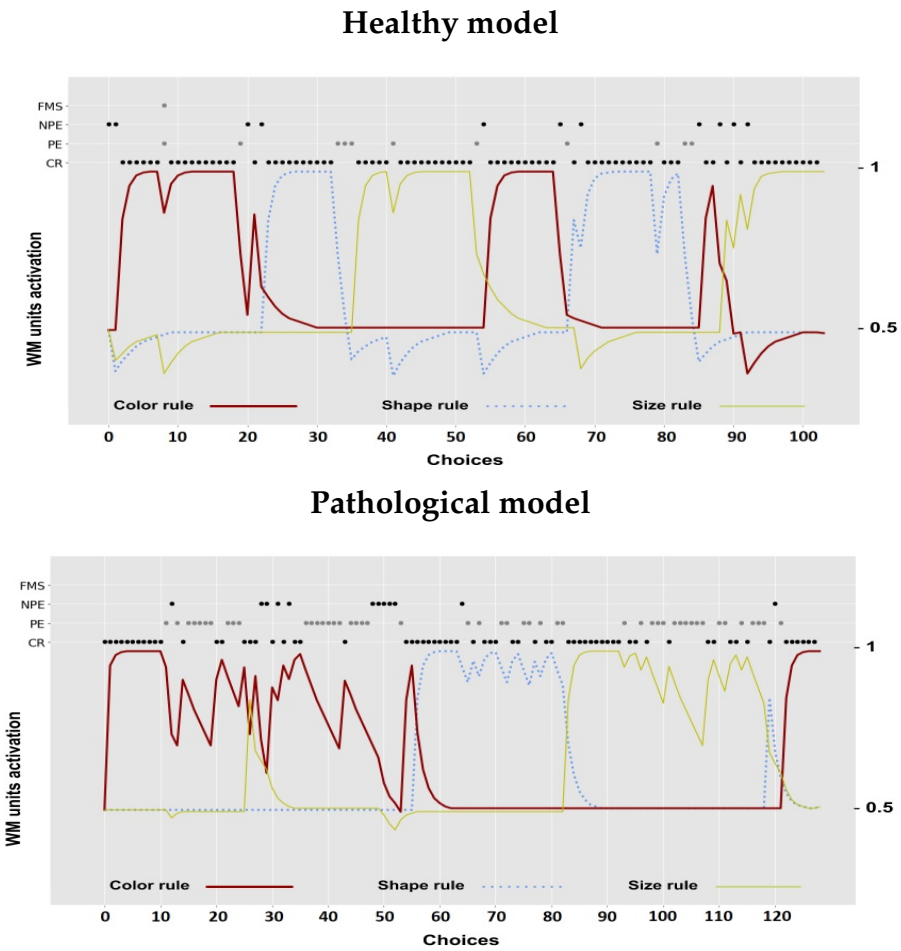


Figure 3.9: Internal functioning of the executive working memory in the healthy model and pathological model. Each line represents the activation of a memory unit encoding a specific matching rule: thick red line: colour-based matching rule; dotted thin blue line: shape-based matching rule; and continuous yellow line: size-based matching rule. The dots at the tops of the graphs indicate single instances of correct responses (CR) or errors (PE, NPE, or FMS errors).

suggest the failure of an inferential reasoning process. A stable horizontal curve coupled with many errors corresponds to a strong perseverance tendency (e.g., see the graph for the EPM). Conversely, a graph with several increases and decreases in different curves indicates an erratic behaviour (e.g., see the graph for DM).

The healthy model (Figure 3.9) completed the six WCST categories and performed correct reasoning after category changes, possibly after one or two errors and 'inference by exclusion'. Moreover, in the choice interval of $85 - 95$, the model had many NPEs after choosing the colour rule and receiving positive feedback in trials 86 and 87. In this case, the correct sorting rule was size but the model chose a target card that shared both the colour and size attributes with the deck card.

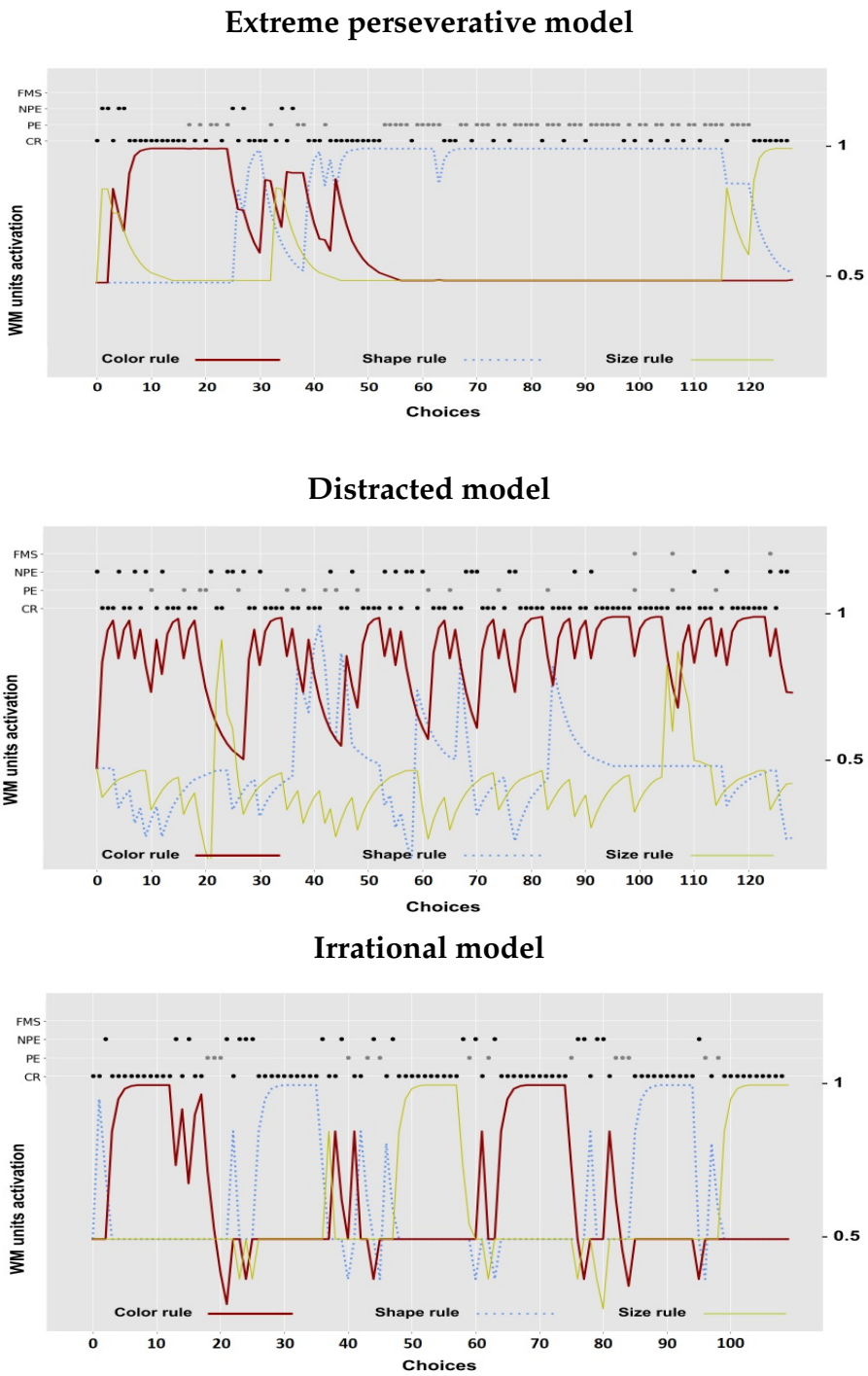


Figure 3.10: Internal functioning of the executive working memory in the models with focused alterations. Each line represents the activation of a memory unit encoding a specific matching rule: thick red line: colour-based matching rule; dotted thin blue line: shape-based matching rule; and continuous yellow line: size-based matching rule. The dots at the top of the graphs indicate instances of correct responses (CR) or errors (PE, NPE, or FMS errors).

The model focused on the colour rule (the red solid line representing the colour priority increased after positive feedback) and the positive feedback led the model to increase the priority of the colour rule. Next, in trial 88, the deck card and target

card shared the colour attribute but not the size attribute (the correct sorting rule was still size), so the model received negative feedback and it lowered the colour priority and chose the correct size rule.

The pathological model (Figure 3.9) produced many prolonged incorrect activations that led to both PEs (e.g., in the choice intervals of 15–25 and 35–45) and NPEs (in the choice intervals of 27–33 and 47–53). Despite the presence of both error types, the model had two long series of PEs (choice intervals of 65–82 and 95–120) and it continued to choose the size rule after changing from size to colour.

The EPM (Figure 3.10) obtained a similar trend to the pathological model but with more prolonged incorrect fixed choices (e.g., see the choice interval of 65–115). Interestingly, the inability to switch the sorting rule after negative feedback caused many small perseverative trends during the inferential reasoning process (e.g., see the choice interval of 30–40).

The DM (Figure 3.10) had a high number of sudden random changes in working-memory activations, which caused many NPEs. Interestingly, the model also produced scattered PEs (see Section 5 for an explanation of this phenomenon). Moreover, due to the erratic behaviour, the model often chose an incorrect rule despite its low priority value, thereby lowering it further (e.g., see the choice interval of 54–59).

The IM (Figure 3.10) produced an almost healthy-like plot, with the fundamental exception that many errors were produced when the model should have changed the matching rule. In this case, the priority values of all the rules immediately dropped to the same baseline, and the model did not keep track of the effects of past actions or prevent the execution of bad choices based on previous feedback. As a consequence, on average, the model had to make more choices to find the correct rule in a random manner, and thus it incurred some PEs and several NPEs.

Perceptual component: internal representations After training , the DBN that implemented the perceptual component could extract the specific attributes of each card with its highest neuron layer while also generating the images corresponding

to these attributes in the input layer by interacting with the top-down manipulator. The model used the latter capacity based on its *generativity* to compare the WCST cards in the selected attribute category (colour, form, and size). To investigate the quality of these representations, I analysed the images reconstructed by the component when single units from the first and second hidden layers in the network were manually activated in isolation (this operation simulated the disinhibition effect of the top-down manipulator when performing the WCST). Figure 3.11 shows the images generated by activating the units in the first hidden layer (graphs on the left) or the units in the second hidden layer (graphs on the right). The figure shows that the images generated by activating single units in the first hidden layer involved different attributes of categories (e.g., mixed colours, forms, or sizes). By contrast, the images obtained by activating single units in the second hidden layer involved disentangled representations of each specific category attribute independently of the other attributes. For example, a unit encoded the 'prototype' of the blue colour independently of the size and shape of the object, and another unit encoded the prototype of the triangle shape independently of the colour and size of the object.

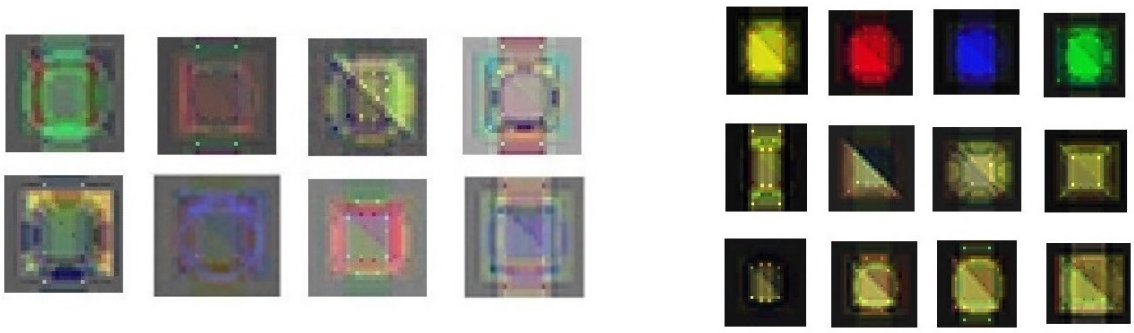


Figure 3.11: *Left*: Images generated by activating a sample of single neurons in the first hidden layer to show how each encodes a mixture of colour, shape, and size attributes. *Right*: Images generated by activating single neurons in the second hidden layer to show how each image encodes a specific disentangled category attribute, which can be seen by considering that the three rows of graphs refer to the three categories (from top to bottom: colour, form, and size) and the four columns refer to different category attributes (colour: yellow, red, blue, and green; form: bar, triangle, circle, and square; size: small, medium small, medium large, and large).

3.1.5 Discussion

Interpretations of the results

Cognitive profiles of the simulated participants The results obtained by the parameter fitting procedure described in Section 3.1.4 (see Figure 3.5 for an overview) showed that the cognitive profiles of healthy young participants were very different compared with those of the three groups of healthy old participants, frontal patients, and Parkinson patients, thereby supporting the idea that the effect of ageing on healthy old participants impairs the executive functions (Dennis & Cabeza, 2012; Sullivan et al., 2001). The 'pathological model' fitted to the frontal patients obtained different values for the μ and ϕ parameters and a similar τ value compared with the 'healthy model' fitted to the healthy young participants. These results suggest that frontal patients: (a) are less flexible at adapting their behaviour after negative feedback (μ); and (b) they have a lower capacity for remembering and reasoning about the appropriate behaviour to undertake based on experience (ϕ). These findings highlight the fact that frontal patients exhibit a mixture of deficits, with a tendency to perseverate in non-adaptive behaviours and poor executive functioning (e.g., see Barceló, 1999).

The model configurations fitted to Parkinson patients and healthy old participants obtained very different profiles. In particular, Parkinson patients exhibited less sensitivity to negative feedback (μ) compared with the paired control group. This difference might have been related to their altered capacity for processing reward and feedback, which is a distinctive feature of the disease caused by the corruption of the dopamine system (Volpato et al., 2016). The comparison also indicated higher distractibility (τ), which might have been related to the lower capacity of Parkinson patients to 'lock-in' on the correct behaviour, and this is another relevant function of the dopamine system (Fiore et al., 2014). The comparison also highlighted an unexpected result where the working-memory forgetting (ϕ) of Parkinson patients was low compared with the related control group (healthy old participants), although it was still higher than that of the healthy young

participants in the study by Heaton et al. (2000). This result could be explained by the effects of Parkinson treatments on the executive role of the working memory, thereby possibly affecting the reasoning-by-exclusion process involved in the performance of the WCST (Fallon et al., 2017).

A second unexpected result was that the working-memory forgetting speed (ϕ) of healthy old participants was similar to that of the frontal patient group members in the study by Heaton et al. (2000). This result has an interesting explanation, which was captured by the model, as follows. The frontal patient group had an average age of 42 ± 14.32 years whereas the healthy old participants group had an average age of 69.74 ± 6.96 years, and thus the latter group was probably affected by age-related weakening of the working memory (Daselaar et al., 2013).

Cognitive processes and behavioural responses The results reported in Section 3.1.4 highlight the interesting relationships between cognitive processes and the behavioural indices scored in the WCST. For example, the positive correlation between CC (indicating global performance) and the error sensitivity (μ) supported the construct validity of the WCST, that is, the test evaluates the capacity to change the categorisation rule after negative feedback. Furthermore, in order to exhibit adequate performance in the WCST, a participant requires an intact working memory storage capacity (negative correlation between CC and the working memory forgetting speed parameter, ϕ) and attention abilities (negative correlation between CC and the distractibility parameter, τ).

The correlations between TE and the behavioural indices supported my considerations regarding the CC index (both the CC and TE indices indicate the global performance in the WCST). In particular, the negative correlation between TE and the error sensitivity parameter (μ), as well as its positive correlation with the distractibility parameter (τ), supported the idea that negative feedback reactivity and attention abilities (operationalised as 'distractibility' in this study) are key processes when solving the WCST.

The negative correlation between PEs and the error sensitivity parameter (μ) sup-

ported the findings of classic studies of the WCST, which associated perseverative rigid behaviours with difficulty in adapting behaviour after negative feedback (Dehaene & Changeux, 1991).

Interestingly, the positive correlations between NPE with distractibility (τ) and the working memory forgetting speed (ϕ) confirmed the previously claimed relationships between this type of error, the lack of attention abilities, and 'reason by exclusion' failures (Dehaene & Changeux, 1991; Barceló & Knight, 2002). Moreover, the negative correlation between NPEs and the error sensitivity parameter (μ) suggested that reasoning by exclusion (the failure of which tends to cause NPEs) strongly depends on the capacity to evaluate external feedback.

Finally, the strong correlation between the FMS errors and distractibility (τ) confirmed that maintaining correct behaviour is highly dependent on the ability to maintain internal focus on the selected categorisation rule. Overall, these results are in agreement with previous findings (Section 3.1.1) and the results obtained by different models (Section 3.1.5).

Brain lesions, cognitive deficits, and behavioural impairments The results presented in Section 3.1.4 and summarised in Figure 3.12 show the possible correspondences between model lesions and brain lesions, and the consequent behavioural impairments.

The EPM characterised by a very low error sensitivity (low μ parameter) had a high number of PEs. This lesion might correspond to malfunctioning of the ventral ACC involved in the motivational processing of errors (Lie et al., 2006). This structure together with medial and ventral cortical and sub-cortical areas regulates negative emotions (Etkin et al., 2011) and the processing of the affective valence of stimuli (Roy et al., 2012).

The DM characterised by high distractibility (high τ parameter) obtained the opposite behavioural profile compared with the EPM, that is, an index imbalance toward NPEs compared with PEs, although with only a minor difference. Moreover, the DM had an increased number of FMS errors, thereby confirming an

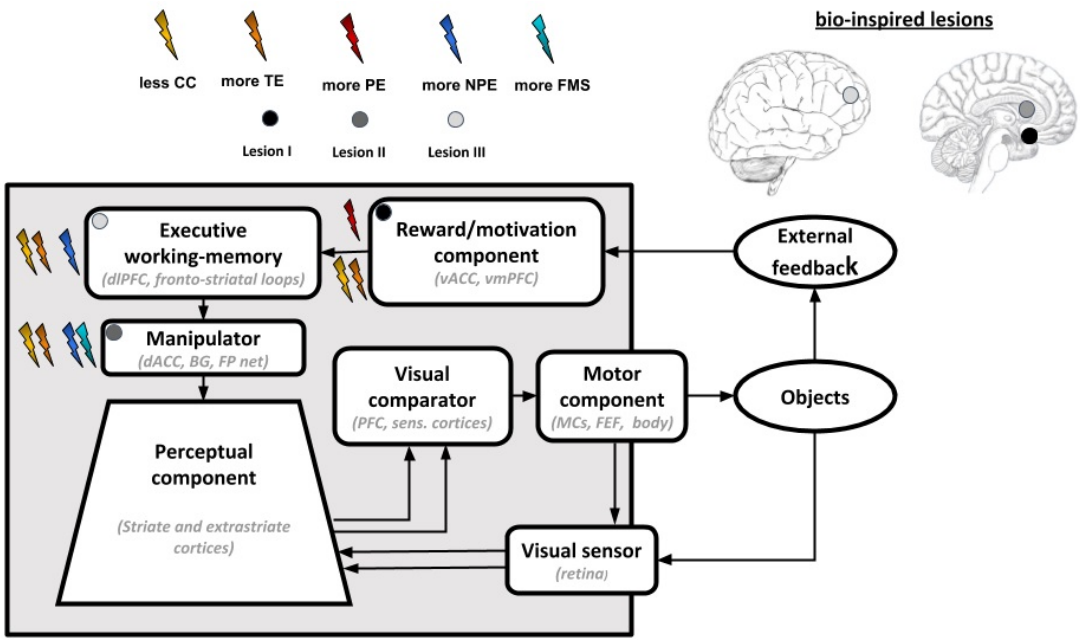


Figure 3.12: Schema showing the architecture of the model and three 'focused lesions' obtained by altering specific parameters, which I applied to obtain three prototypical pathological conditions. Dots with a different intensity of grey represent the three alterations ('lesions') and the coloured bolts denote the decrease in performance/increase in errors that they caused during the performance of the WCST. As an example, lesion I mimicked effects analogous to those produced by a brain lesion in the ventral ACC and ventromedial PFC to impair the motivation system, and it caused a decrease in CC and increases in TEs and PEs.

unstable attention focus. This alteration might correspond to an impairment of the dorsal ACC that interacts with the dorsal and frontal cortices to influence decision making and response selection (Bush et al., 2002; Heilbronner & Hayden, 2016).

Finally, the IM characterised by a high working memory decay speed (high ϕ parameter) had a slight imbalance towards NPEs. This alteration caused the working memory component to have a high forgetting rate for previously chosen rules and it might correspond to an impairment of the brain system that supports executive working memory, particularly the dorsolateral PFC and its loops with basal ganglia (Mannella et al., 2013).

Working memory dynamics and consequent behaviour in the WCST The results presented in Section 3.1.4 showed that different behavioural responses to the WCST corresponded to different executive working memory dynamics. The

rule-based activation of the executive working memory of the model appeared qualitatively similar to those found in the neurons of the dorsolateral PFC of non-human primates performing variants of the WCST (Mansouri et al., 2006; Buschman et al., 2012). These results indicate that executive working memory storage and updating are key processes when executing an adequate internal manipulation of representations, and thus they support flexible behaviour. For example, the healthy model fitted to the young healthy participants obtained overall good performance. However, it still exhibited exploratory behaviours supported by unstable activation of working-memory sub-goals. These behaviours might appear pathological. This phenomenon highlights the fact that the global behaviour of a participant can exhibit occasional cognitive failures.

The pathological model fitted to the data for frontal patients exhibited both perseverative behaviour (many PEs) and reasoning failures (many NPEs), thereby supporting the idea that frontal patients can be affected by different cognitive deficits beyond behavioural rigidity, such as working memory impairments characterised by inferential reasoning failures. Moreover, PEs are also exacerbated by a specific feature of Heaton's version of the WCST where the deck and target cards sometimes have more than one attribute in common, which can produce positive feedback regardless of whether the participant performs sorting based on the wrong matching rule (Dehaene & Changeux, 1991).

The EPM exhibited similar behaviour to the pathological model but it was characterised by a more severe insensitivity to feedback, thereby resulting in a higher number of PEs compared with the previous pathological model. The EPM was also more strongly affected by the feature of Heaton's version of the WCST described above than the pathological model. These dynamics support the relationship between cognitive rigidity involving feedback-independent maintenance of the same specific sorting rule and perseverative behaviour.

The DM exhibited erratic behaviour caused by severe impairment of the decision-making processes. In particular, this model produced a 'stimulus-driven behaviour', which was dissociated from the rule priority values, and it yielded a

response based on one of the random specific attributes suggested by the input (colour, shape, or size). This behaviour resulted in the model frequently choosing a strategy with low desirability at a high level, thereby obtaining the lowest values for its working memory units compared with the other versions of the model. These impaired dynamics highlight the importance of attentional focus during the WCST because its deficit can cause unstable behaviour.

Finally, the IM exhibited healthy behaviour but with a highly impaired capacity for reasoning by exclusion, as shown by the fact that when the rule changed, the model required many attempts to discover the new rule. As found in experiments with human participants, it should be noted that the simulated experimenter represented by a software routine detected the errors but had no access to the decision-making processes of the model. This feature is a potential limitation of this version of the WCST due to many different cognitive factors such as an erratic decision-making process rather than repeated intentional wrong rule selection resulting in PEs. This limitation makes it more difficult to interpret the link between this behavioural index and the underlying cognitive processes.

Main theoretical contributions

The aim of this study was to operationalized and corroborate the three-component theory, a novel theoretical hypothesis that states that flexible cognition depends on the top-down manipulation of internal low-level perceptual representations (see section 2.3 for further details).

The tests showed that the model can perform the WCST. In particular, the core mechanism in the model allows a behavioural rule (goal) selected within the executive working memory to apply a top-down bias on the lower perceptual levels. This bias leads to a representation of the input that reflects the selected rule. The model diverges from previous models of the WCST (see the following section for further details) that directly link the selection of behavioural rules to the selection of actions. Instead, the model selects the manner in which the inputs are

internally represented and these goal-biased representations then trigger suitable actions. Thus, the model represents a new tool for quantitatively studying the proposed hypothesis. This possibility was demonstrated in the present study by validating the model with data from multiple WCST experiments involving healthy young and old participants, as well as frontal and Parkinson patients, which have often been reproduced in isolation using previous models (Table 3.2 and Figure 3.7).

Qualitative analysis of the internal representations of the executive working memory of the model demonstrated the key role that the internal manipulation of representations can play in flexible behaviour. In particular, this manipulation allowed the model to focus on the correct rule (sub-goal), and thus to perceive the cards in a 'rule-biased manner' that was suitable for supporting the correct responses (Figures 3.9 and 3.10). The activations of the executive working memory of the model are compatible with those found in the PFC during the performance of the WCST (Mansouri et al., 2006; Buschman et al., 2012). Furthermore, the presence of perceptual representations with different levels of abstraction within the first and second hidden layers of the perceptual component reflects hierarchical information processing in the perceptual cortices of the brain (Felleman & Van Essen, 1991; Mechelli et al., 2004; Baldassarre et al., 2013a). For example, the neurons in the primary visual cortex extract several low-level visual features from retina images (Rentzeperis et al., 2014) whereas the neurons in the higher-order visual cortices tend to respond to macroscopic aspects of objects (Bracci et al., 2017; Folstein et al., 2015).

Comparisons with other computational models

In this section, I present comparisons of my model with previously proposed computational models for studying the cognitive processes and neural mechanisms that underlie the performance of the WCST. Moreover, I consider other models that have not been used for studying the WCST but that have been employed for investigating executive functions and proposing hypotheses regarding the

cognitive processes and biological mechanisms that support flexible cognition.

Models of WCST Table 3.5 summarises the main features of the computational models used to investigate the WSCT, including my model.

Levine & Prueitt (1989) proposed a model for performing the WCST based on adaptive resonance theory (Carpenter & Grossberg, 1987). This model suggests that categorisation in the brain is based on an interactive relationship between top-down processes (e.g., expectations) and bottom-up processes (sensory information). The model qualitatively reproduces both perseveration and the novelty dependence of frontal patients. These behaviours are linked to the impaired integration of frontal structures that support both cognitive processes (attention to specific rules) and motivational processes (past effects of decision-related rewards and punishments). By contrast, my model supports the idea that a corrupted link between feedback computations (past rewards and punishments) and attention selection (selection of a specific rule) is caused by an impaired rule selection process (see DM in Section 3.1.4). In my model, this corruption produced slightly more PEs and many more NPEs, which were not considered by the authors.

Models of WCST	Functions/Computational elements					Biological constraints	Data fitted	Number of free parameters
	Working Memory	Rule selection	Feedback computation	Sensory-motor processes	Top-down manipulation			
Levine & Prueitt (1989)	✓	✓	✓	✗	✗	✗	✗	1
Dehaene & Changeux (1991)	✓	✓	✓	✓	✗	✗	✗	3
Berdia & Metz (1998)	✓	✓	✓	✗	✗	✓	✓(2)	2
Amos (2000)	✓	✓	✓	✗	✗	✓	✓(6)	4
Kaplan et al. (2006)	✓	✓	✓	✗	✗	✗	✓(2)	2
Bishara et al. (2010); Steinke et al. (2018)	✓	✓	✓	✗	✗	✗	✓(7)	4
Caso & Cooper (2017, 2020)	✓	✓	✓	✓	✗	✓	✗	4
Steinke et al. (2020a,b)	✓	✓	✓	✓	✗	✗	✓(4)	8
This model	✓	✓	✓	✓	✓	✗	✓(4)	3

Table 3.5: Overview of the main features of computational models used to investigate the WCST. 'Biological constraints' indicates whether the model incorporates fine-grained neural details (i.e., bio-constrained neuron models and detailed micro circuit connectivity; the other models, as mine, capture only the interactions between the brain macro-systems underlying the WCST). 'Data fitted' indicates whether the model was used to fit human experimental data (e.g., behavioural indices obtained during the solution of WCST), and the number in brackets indicates how many different data sets were used.

Dehaene & Changeux (1991) proposed a model of the WCST that encompasses the top-down selection of rules and their integration with percepts, and a reward signal to select actions. This model also considers an 'intention layer' linked to the choice of the four target cards. The model reproduces PEs and a worsening of 'single-trial learning' as an index for measuring the length of a successfully completed series, which was not considered in the present study. These two results are based on impaired feedback processing and rule-based memory corruption. This model was not proposed recently but it incorporates various possible explanations of synaptic and molecular processes as the basis of solving the WCST. These processes are simulated at an abstract level, and they are related to reward processing and synaptic plasticity. Overall, this previous study provided an extensive functional analysis of the task, but it mostly focused on the perseverative tendency of patients in the WCST. In contrast to my proposed model, this previous model fails to analyse other types of errors, such as NPEs and FMS errors, thereby preventing the possibility of effectively discerning patient sub-populations, as achieved in my study, particularly determining the heterogeneous deficit profiles related to distractibility and perseveration.

Berdia & Metz (1998) proposed a model that simulates neural noise and synaptic instability based on two parameters comprising 'noise' and 'gain', and they linked them to the poor performance of schizophrenic patients in the WCST. This is one of the first models of the WCST to highlight the idea (as supported by my model) that a decrease in the influence of motivation (rewards/punishments) on behaviour can increase NPEs, thereby explaining poor global performance (e.g., in schizophrenic patients). The model considers attention processes related to the competition between categories, and reproduces PEs and NPEs in normal and schizophrenic participants. However, this model does not consider FMS errors (a sub-set of NPEs), which I included in my model. This omission prevents the investigation of multiple causes of NPEs, and particularly FMS errors. In my study, I found that NPEs can be caused by a reasoning-by-exclusion failure (IM) or by an unstable attention focus (DM), but only the latter type of failure caused a high

number of FMS errors.

Amos (2000) proposed a model of the WCST that reproduces cortico-striatal loops and the related involvement of dopamine. In particular, their model suggests that the frontal cortex stores and selects the behavioural rule to follow, and that the striatum selects the target card based on the input card. By altering the two parameters linked to these key components, the model could fit the global performance (CC and TE) and the PEs of three groups of patients and related control groups. In particular, the model proposes that schizophrenic patients are affected by frontal impairment whereas Parkinson patients are affected by striatum deterioration. The model reproduces the behaviour of many human groups but it does not include non-perseverative and FMS errors, which are important for discerning many types of brain lesions. Furthermore, their model suggests that Parkinson patients are defined by a specific sub-cortical impairment, but my model showed that Parkinson patients can be characterized better by a heterogeneous impairment profile. In particular, Parkinson patients exhibit deficits involving both error sensitivity related to frontal-ventral impairment and distractibility related to alterations of both dorsal regions (e.g., dorsal ACC) and ventral regions (e.g., striatum).

Kaplan et al. (2006) reproduced the WCST with a model that uses a *Hamming network* (a feed-forward neural network for solving pattern recognition problems; Lippmann, 1987) to generate new strategies and a *Hopfield model* (an associative neural network; Hopfield, 1982) for storing them. These networks were used to reproduce both perseverative and failure-to-maintain errors in healthy and prefrontal patients. In this model, it is assumed that the former are caused by rigidity and the latter by attention failures. Similar to other models, this model does not consider NPEs, which I linked to both attention failures and failures of inferential reasoning in the present study. Furthermore, this model does not consider that attentional failures can cause PEs, which was shown by my model.

Bishara et al. (2010) proposed a model for performing the WCST and investigating the cognitive profiles of patients with substance addiction, schizophrenia (Cella

et al., 2014), bipolar disorder (Farreny et al., 2016; Cella et al., 2014), and Parkinson patients (Steinke et al., 2018). The model encompasses an abstract component for computing positive feedback, negative feedback, and a 'choice consistence' (attention focus), as well as two different parameters for regulating the sensitivity to negative and positive feedback based on neuroscientific research that demonstrated a dissociation between the two (Monchi et al., 2004). Despite this evidence, Steinke et al. (2018) applied the model to study Parkinson patients and in agreement with my results, found that a single parameter for modulating the response to negative feedback was sufficient to fit their performance. This previous model can fit data related to a higher number of human groups but it does not consider all of the WCST behavioural indices, as included in my model.

Caso & Cooper (2017, 2020) proposed a computational model of healthy and Parkinson participants performing the WCST. This model aims to operationalise the 'schema theory' proposed by Schmidt (1976) by using a model architecture based on the neuroanatomy of basal ganglia and corticothalamic loops. This model highlights the important function played by basal ganglia as a fundamental 'selection machine' in the brain (Redgrave et al., 1999); this function is also incorporated in my model. Moreover, similar to the model described next, this model stresses the idea that both motor selection (specific target cards) and 'conceptual selection' (sorting rule) influence the performance of participants during the solution of the WCST. It was concluded by Caso & Cooper (2020) that Parkinson patients exhibit a perseverative profile with a strong memory of past feedback, which corresponds to the lower 'memory decay' in my study. In addition, my model indicates that Parkinson patients exhibit high distractibility with respect to the correct strategy to follow, thereby supporting the idea that Parkinson patients exhibit a mixed impaired cognitive profile.

Steinke et al. (2020a) proposed a model of the WCST that depends on the concept of two-level reinforcement learning. In particular, the model suggests that the trial-by-trial behaviour of participants is supported by model-based learning involving a decision-making process based on the sorting rule to choose and

model-free learning based on the motor response executed after feedback (i.e., one of four target cards). The model was used to fit data related to healthy young participants and to show the existence of a perseverative tendency caused by response avoidance after negative feedback. This model was further validated by Steinke et al. (2020b) who fitted two groups of Parkinson patients ('on' and 'off' medication) and a matched healthy control group. The model demonstrated that Parkinson patients had lower sensory-motor competencies (model-free learning processes) and cognitive deficits (model-based learning processes), and that the medication caused further cognitive symptoms such as cognitive inflexibility and attentional failures. My model is comparable to the 'model-based learning' model version (only rule-based action) and it does not encompass a model-free learning component (card-based action). This previous model supported a card-specific effect but it was shown that the model-based choices could have a greater weight and fit better to the behaviour of some participants. Moreover, it was shown that the model with both model-based and model-free learning modalities produced the best fitted results, but it also had high complexity (eight or seven parameters). My model can account for NPEs and PEs, and it can fit a comparable number of human populations based on only three free parameters.

Table 3.5 summarises the features of the models considered. Some of these models have objectives that go beyond the investigation of the WCST. For example, Caso & Cooper (2017, 2020) aimed to test schema theory, which was then operationalised in a model tested with the WCST. Similar to this approach, my principal objective was to investigate the three-component hypothesis based on the mechanisms that underlie flexible cognition and to support the theory by operationalising it in a model validated with WCST data.

I have presented qualitative comparisons of my model and other models, but I might perform quantitative comparisons to obtain more informative outcomes in future research, in a similar manner to that conducted by Steinke et al. (2020a). My qualitative comparison mainly indicates that none of the previously proposed models is supported by the manipulation of perceptual representations within

low-level perceptual areas, and that category-based input representations might play relevant roles in the performance of the WCST. Moreover, some models involve sensory-motor components but in contrast to my model, none of them includes a visual search process that works together with a top-down manipulation mechanism to support flexible behaviour. In particular, most of the models assume the existence of hardwired semi-localistic representations of input patterns (e.g., representations based on one-hot vectors). By contrast, my model generates input representations based on a visual abstraction process applied to the raw visual input patterns. According to this analysis, most models of the WCST are based on a direct sequence of processes, as highlighted in Figure 2.3, which comprises 'rule decision - selection of sub-goal - action performance - feedback computation'. By contrast, my hypothesis and model assume that the high-level decision-making processes related to the sorting rule to follow have a 'backward effect' on the internal low-level representations of stimuli, and these representations then affect action selection.

Models of executive functions Table 3.6 summarises the features of some models proposed for investigating executive functioning and category learning in healthy and pathological participants, which are relevant to the issues investigated in this study.

Ashby et al. (1998) proposed a relevant model called 'COVIS' that represents both a theoretical framework and a computational implementation of neural structures that support category learning processes. This model is based on the hybridisation of symbolic and sub-symbolic mechanisms. In particular, the model is based on the idea that the brain performs category learning through: (a) a procedural implicit system that supports automatic action execution by mostly involving subcortical structures such as basal ganglia; and (b) a logical explicit system that supports rule-based behaviour by mostly involving cortical areas. This model assumes that the manipulation of response locations interferes with the procedural implicit system but not with the rule-based decision-making process.

It was concluded that the WCST is mostly supported by a rule-based information-processing system, which is an idea that is also implemented in my model and other recently proposed models. The model was validated with WCST data collected from healthy participants and Parkinson patients Hélie et al. (2012), and it highlighted the role of dopamine shortages in the executive deficits exhibited by these patients.

Monchi et al. (2000) proposed a biologically plausible model of working memory activation during the solution of two tasks comprising the delayed response task and WCST. In particular, the architecture of the model is based on biologically plausible neurons and it emulates the brain system formed by basal ganglia thalamocortical loops and working memory. The model was also lesioned to investigate working memory deficits in Parkinson and schizophrenia patients, and it predicted that the working memory deficits in Parkinson patients are caused by impaired disinhibition affecting the encoding and storage capacities of specific features. However, the model predicted that the working memory deficits in schizophrenia patients are related to the capacity for selecting specific features to store. Interestingly, my results support the possibility of rule selection impairment in Parkinson patients. In addition, I found that in addition to perseverative behaviour, incorrect selections can be caused by a working memory impairment (overlaps between rule representations) or by an altered top-down selection process that must evaluate the priorities of rules, where only the latter process is impaired in Parkinson patients.

Gilbert & Shallice (2002) propose a simple model of inhibitory control when performing the Stroop task. In particular, the model has two channels comprising a verbal channel and a colour channel, which compete to guide the behavioural response. Furthermore, this model has a 'task demand component' that stores the task requests ('name of the letters in input' or 'name of the colour in input') and a 'top-down control' component that indicates which task to follow. The model assumes that the switching costs between the two task demands are linked to top-down control and not only to automatic processes for response/conflict resolution. Interestingly, this model shares some features with my model but in a

simpler form. In particular, similar to my model, this model performs top-down manipulation of the input representations by executing top-down selection of the input visual features to focus on the letters or the colours, and by implementing an 'intra-category competition' between the observed attributes to activate a specific attribute (e.g., red, green, or blue colours). These similarities with my model demonstrate that cognitive flexibility requires computational components linked to executive functions, particularly inhibitory control and working memory. Moreover, this model supports the idea that executive functions are based on the internal manipulation of representations.

Rougier et al. (2005) proposed a model that was used to reproduce the results of the WCST and it shares features with the other biologically grounded models. The model was implemented within the 'Leabra' framework (O'Reilly & Munakata, 2000) and it comprised various neural maps corresponding to the input layers, parietal cortices, prefrontal cortices, and motor output layers. The model was tested with the WCST and it reproduced the production of PEs when the PFC layer was lesioned, but other types of errors were not considered. In addition to the WCST, this model supports the idea that flexible behaviour is associated with the emergence of distributed rule-like representations, as also shown in the present study.

Another related model (the 'PBWM model') proposed by O'Reilly & Frank (2006) and updated by Hazy et al. (2007) and Kriete et al. (2013) replicated various functions of working memory by using an actor–critic model architecture (Sutton et al., 1998) to reproduce the functions and macro-anatomy of the basal ganglia. In particular, this model was used to demonstrate the fundamental role of 'gating units', which are possibly used by basal ganglia to perform the uploading, storage, and download of information in working memory. This function is abstracted in my model by the winner-take-all competition involving the working memory units, and the basal ganglia disinhibition mechanism is used by the manipulator to allow the working memory to select lower-level perceptual representations.

Finally, Rigotti et al. (2010) proposed a recurrent neural network that allows rule

selectivity in a version of WCST involving only 'form' and 'colour' categories. The model reproduced the internal neuronal dynamics involving mixed selectivity neurons, thereby suggesting that randomly connected neurons spontaneously exhibit mixed selectivity. The model did not assume an architecture for performing the WCST but instead it focused on the internal processes needed to solve context-dependent tasks with the aim of investigating how integrated neural networks can support the selection and storage of behavioural strategies.

Models of executive functions	**Functions/Computational elements**					**Biological constraints**
	Working memory	**Rule selection**	**Feedback computation**	**Sensory-motor processes**	**Top-down manipulation**	
Ashby et al. (1998)	✓	✓	✓	✗	✗	✓
Monchi et al. (2000)	✓	✓	✓	✗	✗	✓
Gilbert & Shallice (2002)	✓	✓	✓	✗	✓	✗
Rougier et al. (2005)	✓	✓	✓	✓	✗	✓
O'Reilly & Frank (2006) Hazy et al. (2007) Kriete et al. (2013)	✓	✓	✓	✓	✗	✓
Rigotti et al. (2010)	✓	✓	✓	✗	✗	✓
This model	✓	✓	✓	✓	✓	✗

Table 3.6: Overview of computational models proposed to investigate executive functions and brain networks relevant to the issues investigated in the present study.

Limitations and future work

Despite the contributions highlighted in the previous sections, the current model has some limitations that could be addressed in future work. In terms of validating the model with empirical data, a future study might aim to investigate the complexity of the model with respect to its parameters. In particular, as shown by Steinke et al. (2020a), I could produce many versions of the model with different numbers of free parameters and compare them based on various indices that consider both the fitting accuracy and complexity of the model, such as the *Bayesian information criterion* (Schwarz et al., 1978).

A specific aspect of the architecture of the model that needs to be improved is the information flow between its components. A general strategy to address this

issues could involve the use of deep neural networks (Goodfellow et al., 2017), as applied to the visual component (for details of this strategy, see the studies by Naselaris et al., 2018). Another possibly complementary strategy could involve grounding the information flows in the model by using a wholly neural dynamical system that mimics the macro-structure of the relevant brain components and that does not require a hard-coded algorithm to control the information flows between the components of the architecture (e.g., as applied by Baldassarre et al., 2013b and by Mannella & Baldassarre, 2015a).

Among the components of the model, a first limitation involves the simplicity of the executive working memory component, which comprises a few neural units for encoding the possible matching rules. This component might be improved by using mechanisms employed in other models of working memory, thereby enhancing the biological plausibility of the model (e.g., O'Reilly & Frank, 2006; Rigotti et al., 2010), or those used in deep neural networks (e.g., Hochreiter & Schmidhuber, 1997). Furthermore, the model can support an inferential process (i.e., reasoning-by-exclusion) but it cannot execute 'one-shot second-order inference'. For example, in the cases where (a) the model focuses on the colour feature, (b) the deck card and the target card share the colour and shape attributes, and (c) the model receives negative feedback, it decreases the priority value of the rule on which it is focusing (colour) but not that for the shape feature that is also not correct (in this case, the unique possible correct rule is the attribute not shared between the two cards, i.e., the number). In the future, this limitation could be addressed by implementing an internal reasoning process that considers both the specific feature on which the model focuses and by a further internal simulation of the potential feedback that would be obtained by alternative responses.

Another component of the model that could be enhanced is the overt attentional system, which currently depends only on a bottom-up attention process that allows the model to explore all stimuli in a stereotyped sequential manner. This approach is sufficient to study the WCST but this component might be improved by adding a top-down attention process for the goal-directed exploration of the

elements in the environment, thereby improving the sensory–motor processes in the model and allowing it to perform other tasks (e.g., see Ognibene & Baldassarre, 2015; Sperati & Baldassarre, 2018).

Another important aspect of the model that should be enhanced is the process employed to acquire category and attribute representations. The current model uses a supervised learning algorithm, where the supervision is conducted by an unspecified external mechanism, e.g., other agents. Social learning might be important for the acquisition of categories, but I consider that most category learning by humans is derived from direct experience in the environment. This theory is at the basis of the third model (section 3.3) that employs reinforcement learning algorithms (Sutton et al., 1998; Caligiore et al., 2019a) to support an autonomous learning of categorical representations on the basis of and performances-dependent environment feedback. Future investigations could attest a progression between the third model and this model, as a continuum in the human cognitive development from infants representation learning processes to adults representation manipulation.

A further improvement involves the key mechanism guiding the internal manipulation of representations. The three-component theories expects that the representation manipulation can involve both perceptual representations and hihg-order representations. However, the manipulation mechanism of this model operates only on the highest level of the DBN used to implement the visual hierarchy. Therefore an important enhancement of the model can involve a self-directed manipulator that can operate at multiple levels of abstraction.

The second computational model of this research project (section 3.2) represents an improvement in this direction, indeed shows both a perceptual manipulation and an inner-speech component that is able to execute a sefl-directed high-order manipulation of representations.

3.1.6 Conclusions

In this study, I corroborated the three-component hypothesis of flexible cognition, as manifested by human participants performing the WCST. In particular, the hypothesis was corroborated and operationalised by realising a computational model. This model depends on three main processes, which I assume are supported by specific brain systems. The first process, which involves the executive working memory that depends on the brain PFC and ventromedial basal ganglia, stores goals and behavioural rules. The second process, which involves perceptual working memories that depend on hierarchies of perceptual cortical systems, can extract and retain information at different levels of abstraction and generate lower representations based on the activation of patterns encoded in the higher levels. The third process, which involves the internal manipulation of perceptual representations by the brain system comprising the frontoparietal cortices and the underlying dorsomedial basal ganglia-thalamus system, selects the representations in the perceptual working memory based on the activations in the executive working memory. I validated the model by showing that it can reproduce and account for a large set of behavioural indices and data related to healthy and pathological participants in the WCST at the state-of-the-art level. These results corroborate and further articulate the three-component hypothesis, for which the internal manipulation of representations is a core process underlying goal-directed flexible cognition.

3.2 Model 2. Inner speech, an auxiliary process that improves internal manipulation and flexible cognition

Here I introduce the second computational study that corroborates the extended form of three-component hypothesis, focusing on the the role of inner-speech as

a second form of internal representations manipulation. Interestingly, updating the previous model by adding a new inner-speech component results in different model dynamics, outputs, and interactions. Considering these differences and a different focus of this study (inner-speech and representations manipulation), I have chosen to dedicate a specific section to this model. The following sections introduce the task experimental conditions, computational components of the model and the results. At last I propose the discussions and conclusions about this study.

3.2.1 Wisconsin Card Sorting Test and different experimental conditions

The model solves the WCST (see Section 3.1.1 for the presentation of this task) and its performance was compared with human performance during the solution of WCST coupled to the shadowing protocols, in particular the model was compared with the dataset previously published in Baldo et al. (2005). Human participants were psychology college students (average age: 20 years). The protocol used included three experimental conditions (Figure 3.13, bottom): control condition - participants solved the basic WCST; motor tapping - participants solved the WCST while executing a finger tapping task following a rhythmic sound; verbal shadowing - participants solved the WCST while vocally repeating the sound 'Na, Na, ...' following a rhythmic sound. The authors report that the participants who solved the WCST during the verbal shadowing protocol exhibit a behavioural impairment (i.e., an increase of behavioural errors, see below) compared with those that solved the WCST with no interfering protocol (control condition). The authors also reported an impairment, although lighter, in the participants that solved the WCST during the motor-tapping protocol. The authors interpreted this outcome by suggesting that inner-speech, together with other processes such as attention or WM, is a cognitive support for problem solving processes. Furthermore, also based on cross-cultural data, they hypothesised that there are individual differences with respect to inner-speech use during problem solving processes.

The model scoring is the same of the first computational study, namely it follows the official documentation (Heaton et al., 2000) proposing five principal behavioural indices that give a full behavioural and cognitive profile of the test performance: CC - *Completed Categories*, indicating the number, out of five, of the performed non-interrupted ten-card sequences of correct sorting; TE - *Total Errors*, indicating the global performance/deficit; PE - *Perseverative Errors*, indicating a perseverative behaviour; NPE - *Non Perseverative Errors*, indicating an attentional failure or an incorrect inferential reasoning; FMS - *Failure to Maintain Set*, indicating a distracted behaviour.

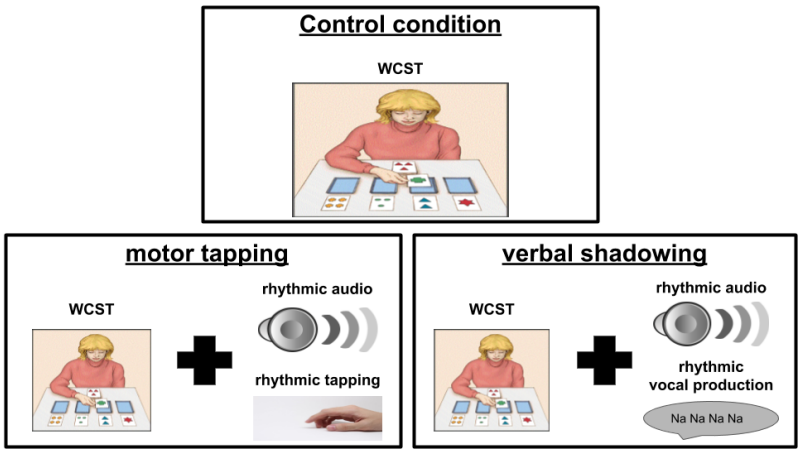


Figure 3.13: Experimental protocols used to test the model, involving the basic WCST (control), and a WCST where the participant has to perform a rhythmic tapping following a rhythmic audio, and a critical analogous verbal-shadowing condition affecting inner speech.

3.2.2 Overview of functioning of the components: key components and dynamics

This model builds on the first model, so it is formed by the same neuro-inspired components that support a number of functions needed to support goal-directed behaviour (Figure 3.14). To summarise, the components that this model inherits from the first one are: (a) Visual sensor: this component extracts visual information from deck cards and target cards, analogously to the eye retina; (b) Hierarchical perceptual component: this component extracts input visual features at increasing levels of abstraction, analogously to the visual brain system (Konen & Kastner,

2008); if activated by top-down mechanisms, this component can also re-generate relevant aspects of percepts, e.g. based on imagination mechanisms (Kosslyn, 1999); (c) Abstract working-memory: this component stores the task sub-goals (sorting rules) chosen by the model, a function that in the brain is mostly supported by frontal cortices (Barraclough et al., 2004; Diamond, 2013); (d) Motivational system: this component uses the external feedback to update the information in working memory, a function that in the brain is mostly supported by ventral basal ganglia (Gläscher et al., 2010; Mannella et al., 2013); (e) Selector: this component chooses a sorting rule and biases the perceptual system (manipulation of internal representations), a function analogously to the top-down control that the fronto-parietal cortex, aided by basal ganglia, exerts on lower-level internal perceptual representations (Redgrave et al., 1999; Gazzaley & Nobre, 2012); (f) Comparator: this component executes visual matching of the deck and target cards, based on the comparison of low-level perceptual representations of the cards by simulating the attentional/imagination processes; in the brain these processes might rely on a distributed network involving the frontal and temporal-occipital cortices (Perani et al., 1999; Kosslyn, 1999); (g) Motor system: this component controls saccades and actions displacing the deck cards close to the chosen target card.

In addition, I added a new key further component to the model, the *inner-speech component* inspired by the brain networks that integrate linguistic and emotional information (Kotz et al., 2006; Sidtis et al., 2018). First, the component transmits information on the relevance of rules to the working memory, in particular information on the sub-goals (identity of the rule) whose priority should be changed, and the positive/negative valence, and intensity, of such change. Moreover, the component implements a phonological-loop storing the current rule independently of its possible pragmatic use.

For an extended description of the model computational components see section 3.1.2 and section 3.1.3. The model architecture is summarised in Figure 3.14, while the following paragraph describes the computational details of inner speech component.

This component is formed by a multi-layer perceptron (MLP). It receives one-to-one connections from the selector units and sends one-to-one connections to the WM units. This process is in particular implemented as follows:

$$m_t = m_{t-1} + \lambda \cdot L_t \tag{3.4}$$

where m_t is the new activation of a WM rule unit, m_{t-1} is the current activation of the WM unit, λ represents the strengths of the one-to-one connection weights linking the language component output-layer units to the WM units, L_t is the current activation of the language component output layer caused by the previous selector units' activation (this time mismatch implies that the component implements a phonological memory). The MLP architecture is formed by 4 input units, 10 sigmoid hidden units, and 3 output linear units.

The input-layer 4 units encode: (a) the selector winner-takes-all *one-hot vector* activation; (b) the binary incorrect/correct match feedback encoded with respectively 0/1. The MLP was trained to activate the output-layer 3 units as follow: the unit corresponding to the selected rule learned to produce a $-1/+1$ value based on the match/mismatch feedback; the other two units activated with 0. For example, if the model chooses the colour rule and receives a positive feedback, the input is $[1,0,0,1]$ and the desired output is $[1,0,0]$; conversely, if the model chooses the colour rule and receives a negative feedback the input is $[1,0,0,0]$ and the desired output is $[-1,0,0]$. The language component is activated two times to simulate: (a) the phonological-loop working memory; (b) the feedback-dependent verbal update of the main working memory. In the first activation, the component input layer is activated by the one-hot code of the selector while its feedback unit is activated with 1 (meaning 'maintenance of the current rule'). In the second activation, the component input layer is activated by the selector activation, but in this case the feedback unit value is activated on the basis of the external feedback (0/1), obtained after the action execution (displacement of the card). The contribution of language to the working memory is regulated by a coefficient λ that ranges in $[0,1]$ and represents the strengths of the one-to-one connection weights linking

the language component output layer to the main working memory units. The coefficient λ is the fourth and last important parameter regulating the functioning of the model and investigated in the simulations. The language MLP component is trained before the experiments illustrated in the main text with a supervised learning algorithm (McClelland et al., 1986). In particular, the system is trained with six different input patterns and six different corresponding output patterns encoding respectively the three possible rules and the binary valence with which to activate the units of the main working memory. The learning rate was set to 0.01 and the network was trained till convergence.

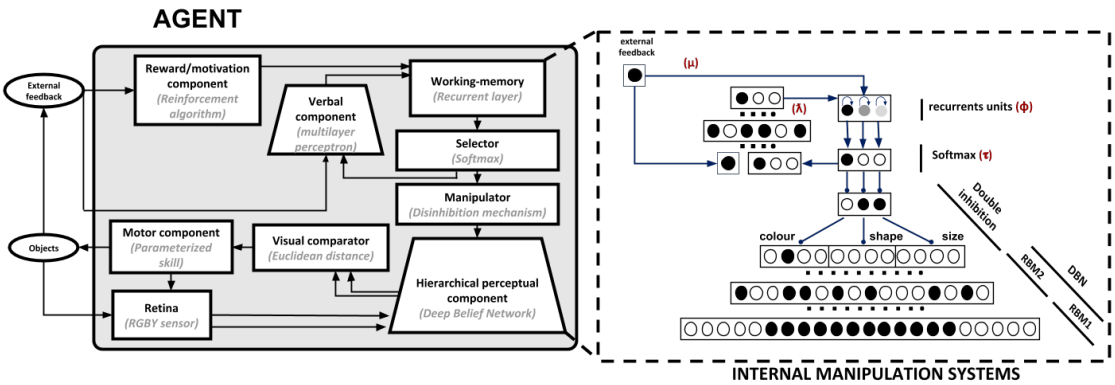


Figure 3.14: Architecture of the model. Left: components of the model. Right: zoom on the neural-network components of the model performing the internal manipulation of representations aided by the language component. The red symbols near the components identify model parameters important for specific cognitive functions (see text for details).

3.2.3 Results

In this section I follow the same organisation of section 3.1.4 of first model. In particular, I first present the results of the comparison of the model behaviour with the human behaviour aimed to investigate the underlying cognitive processes. I then study the model functioning when different aspects of its components are lesioned. Finally, I investigate the internal dynamics of the model with a focus on the role of the language component. Differently from the first study, here I mostly focus on a specific process (inner speech) and its dynamics compared to the validation of whole model. For this motive the section 'validation' is shorter

then the sections that investigate the internal functioning (e.g. Lesions).

Validation of the model with human data

Model configurations that best fit the data from humans I used a statistical search method based on the minimisation of the mean square error (MSE) to find the model parameters that best fit the human data in the three conditions of the WCST, namely the control, motor tapping, and verbal shadowing conditions. The parameters can be interpreted as the relative weights of the simulated cognitive traits of the model (negative feedback sensitivity, memory forgetting, distractability, language contribution), and hence of the modelled human participants. Table 3.7 shows the values of the parameters found with the statistical procedure, now examined in detail.

	Error sensitivity (μ)	**Forgetting speed** (ϕ)	**Distract-ibility** (τ)	**Language contribution** (λ)
Control	0.49	0.97	0.10	0.81
Motor tap.	0.17	0.09	0.12	0.23
Verbal shad.	0.14	0.14	0.13	0.14

Table 3.7: Values of the parameters of the models that produce the best fit of the data on the WCST indices, for the control and experimental groups, reported in Baldo et al. (2005).

I first focus on the parameter λ representing the level of involvement of language processes in the solution of the task. The table shows that the contribution of language is higher in the control condition ($\lambda = 0.81$) than in the motor tapping condition ($\lambda = 0.23$) and verbal shadowing condition ($\lambda = 0.14$). The lower value in the shadowing condition corroborates the model as it indicates that in such condition the model relies on resources other than language to solve the task. The lower value in the tapping condition was instead partially unexpected because this condition should not cause a decrease of the inner-speech contribution. However, since the motor tapping condition involves an auditory process needed to follow the external rhythmic sound, I propose that this interferes with the linguistic contribution to the memory processes as involving the same integrated

phonological processes. Thus the participants rely less on language and more on visual working memory and imagery relying on the non-linguistic processes of the model. However, these alternative solutions represent sub-optimal solutions for humans, used to rely on inner speech, and thus lead to a lower performance with respect to controls. This result can be be considered as a prediction of the model, possibly testable in future empirical experiments.

The control group has a very high ϕ value ($\phi = 0.97$). This suggests a compensating interaction between inner speech, error sensitivity and working-memory information decay. In particular, in case of a repeated strong bias toward a specific rule caused by a high error sensitivity ($\mu = 0.49$), a low distractibility ($\tau = 0.10$), and a high language contribution ($\lambda = 0.81$), a large decay of working memory contents does not prevent a an effective decision making.

The control and experimental groups show a higher distractibility value (τ) and a lower error sensitivity (μ). These results can be explained by an increased cognitive load in these conditions that can cause inefficient decision-making and error detection processes.

Finally, the experimental groups appear less different from each other compared to the control group, corroborating the idea that both experimental conditions cause a performance decrease because both interfere with inner speech.

Comparison between the behaviour of the model and of human groups Figure 3.15 allows a comparison of the behavioural indexes of the model versions with those of the target human groups (control, motor tapping, and verbal shadowing conditions) The comparison shows that there is no statistical difference between them (p-values of t-tests, double tail, $p > 0.05$), thus indicating that the model is very effective in reproducing the behaviour of all the human groups.

Since the model had such a good fit, the role played by the different cognitive processes within the model, quantified by the size of the respective parameters, should reflect an analogous role played in the real participants. Analogously, the *differences* between the different model versions, fitting the behavioural indexes of

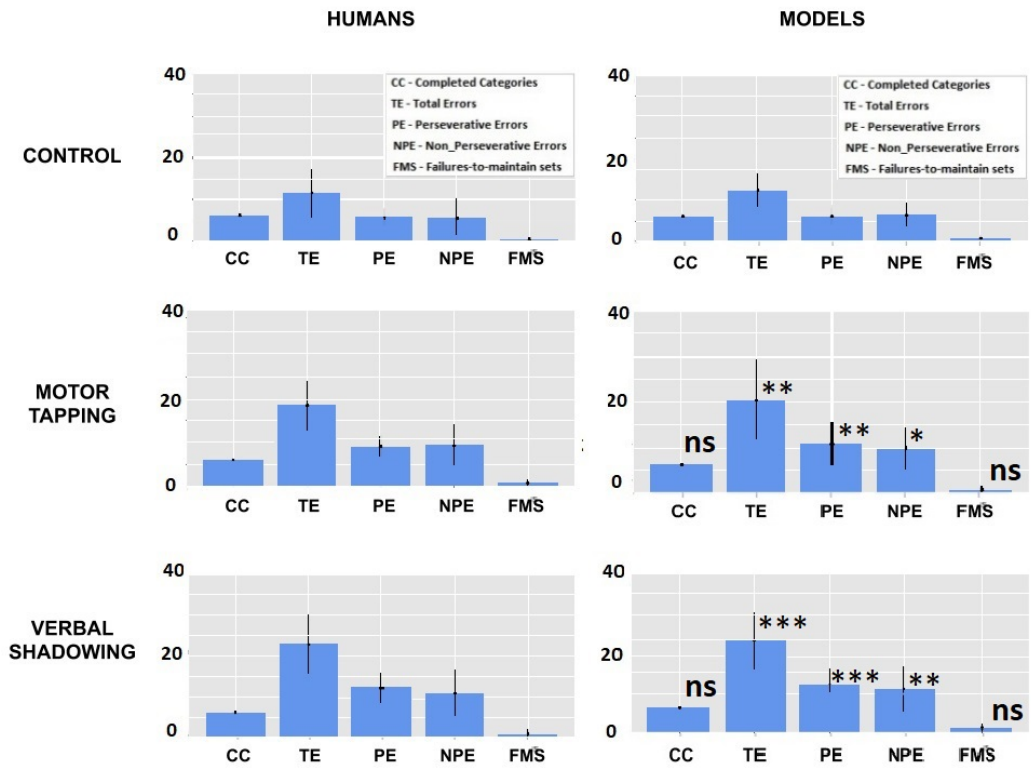


Figure 3.15: Comparison between human groups (left graphs) and models (right graphs) in the three conditions (rows of graphs) for each behavioural index. The significance asterisks in the model graphs are related to the comparison between each of the motor tapping and verbal shadowing models with the control model: ns = non statistically significant, $p > 0.05$; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

the real participants in the different conditions, should reveal the different weight of the cognitive processes in their solution of the WCST.

Figure 3.15 shows a comparison of the behavioural indexes between the different model versions (control model, motor-tapping model, and verbal-shadowing model). While all models did not substantially differ in terms of CC and FMS, the error indexes were higher in the motor tapping model, and even higher in the verbal shadowing model, with respect to the control model. In particular, the motor tapping model exhibited a higher number of total errors with respect to the control group (18.29 ± 5.96 vs. 12.35 ± 4.14; $p < 0.01$), with a balanced profile of PE/NPE errors. The verbal shadowing group showed an even higher number of total errors (22.88 ± 7.34 vs. 12.35 ± 4.14; $P < 0.001$), with an analogous balance of PE/NPE. These results indicate that both motor tapping and verbal shadowing cause a general decrease in performance (stronger for verbal shadowing) due

to both an increased perseverative behaviour (higher PE) and attentional failure (higher NPE).

Study of the model internal functioning

Relation between the model key parameters and the WCST behavioural indices Here I present the correlation results (Pearson's coefficients) between the model parameters, representing the strengths of its cognitive processes, and the behavioural indices scored in the WCST (Table 3.8).

Indices	**Parameters**			
	μ	ϕ	τ	λ
CC	+ 0.00	***- 0.34***	***- 0.72***	***+0.32***
TE	*- 0.05*	***+ 0.40***	***+ 0.70***	***- 0.37***
PE	*- 0.08*	***+ 0.40***	***+ 0.65***	***- 0.40***
NPE	- 0.03	***+ 0.40***	***+ 0.72***	***- 0.35***
FMS	+ 0.01	*+ 0.04*	***+0.86***	*- 0.04*

Table 3.8: Pearson's correlations between key parameters (μ, ϕ, τ, λ) and WCST indices. The table highlights in **bold** the correlation indexes above $|0.3|$, and in *Italics* those that are statistically significant.

The μ parameter (error sensitivity) did not show a strong correlation with anyone of behavioural indexes. However, it showed statistically significant ($p < 0.05$) negative correlations with PE ($r = -0.08$) and TE ($r = -0.05$). These correlations confirm the role of this parameter (error sensitivity) for cognitive flexibility and consequently the occurrence of perseverative errors. The ϕ parameter (forgetting speed) negatively correlated with CC ($r = -0.34$) and showed a moderate positive correlation ($r = +0.40$) with all types of errors with the exception of FMS with which it showed a low but significant correlation. These results suggest that memory decay influences the global cognitive performance of the model with no specificity for errors. The τ parameter (distractibility) had a similar correlation profile, but showed stronger correlations (mostly above a 0.7 value), in particular a strong positive correlation with FMS ($r = +0.86$). This confirms the important effect that distractibility has on FMS errors. The λ parameter (language contribution) shows a positive correlation with CC ($r = +0.32$), and a moderate negative

correlation with all errors with the exception of FMS (non-significant correlation). This suggests that inner speech contributes to global performance similarly to the other processes of attention and working-memory.

I carried out an analysis of the simulations where the role of language was negligible ($\lambda < 0.05$; sample size: $n = 175$; here all correlations between λ and the behavioural indices became non statistically significant). Table 3.9 shows the resulting correlations. This analysis highlights the cross contribution of inner speech to different processes. Indeed, if the contribution of language is strongly reduced and the correlation of the different parameters increases, this means that language can assume a vicarious role with respect to the processes corresponding to those parameters.

The correlation coefficients between μ (error sensitivity) and most behavioural indices became stronger, in particular with PE (from $r = -0.09$ to $r = -0.22$). This indicates that with a weaker language contribution a lower sensitivity to errors increases the model rigidity (perserverance errors); this highlights the important role of language in strengthening the effect of feedback. The positive correlation with FMS ($r = 0.17$) reached significance. This means that μ can polarise the relevance of the different rules and sharpen their selection, and thus reduce the possibility of mistakenly changing the correct rule: this result indicates that language can play an analogous function. The parameter ϕ (forgetting speed) showed a higher correlation with all behavioural indices and in particular the correlation with NPE became stronger (from $r = +0.40$ to $r = +0.72$). The stronger effect of forgetting on NPE means that language can play a compensatory role by biasing the activation of specific rules in working memory. The correlation between ϕ and FMS, which originally had a negligible statistical significance ($r = +0.04$, $p < 0.05$), became stronger and more statistically significant ($r = -0.30$, $p < 0.001$). The implication of this is again that language can strengthen the bias of selecting the correct rule. Conversely, the correlations between τ (distractibility) and behavioural indices passed from strong to moderate values, in particular the correlation with FMS decreased (from $r = +0.86$ to $r = +0.34$). The cause of this,

less clear, is possibly that with a reduced role of language the processes associated with μ and ϕ can more strongly contribute to generate the different errors and so distractability becomes less important.

Indices	**Parameters**			
	μ	ϕ	τ	λ
CC	0.07	***- 0.63***	***- 0.68***	0.05
TE	- 0.10	***0.70***	***0.60***	- 0.09
PE	*- 0.22*	***0.63***	***0.52***	- 0.12
NPE	- 0.03	***0.72***	***0.63***	- 0.08
FMS	*0.17*	***0.30***	***0.34***	- 0.04

Table 3.9: Pearson's correlations between key parameters (μ, ϕ, τ, λ) and WCST indices in the case of a low language contribution ($\lambda < 0.05$). **Bold** indicates correlations above $|0.3|$ and *Italics* the statistically significant ones ($p < 0.05$).

Effects of focused alterations of the model on the WCST behavioural indices

This section presents the study of the effects of seven possible alterations ('lesions') of the parameters of the control model. Table 3.10 shows the specific parameters used to produce each lesioned model. Each of the first three lesions involved a critical parameter regulating a main process of the model, namely error sensitivity, forgetting speed, and distractability. These lesions were supposed to lead to respectively an 'extreme perseverative model' (EPM), a 'distracted model' (DM), and a 'irrational model (IM). I did these lesions to study the compensatory role of language with respect to the functions involved by the parameters. The last four lesions involved the verbal component (inner speech). In particular, the first verbal-lesion model (VLM1) involved a reduced contribution of language in case of external negative feedback. The second verbal-lesion model (VLM2) involved a reduced contribution of language in case of external positive feedback. The third verbal-lesion model (VLM3) involved a low language contribution during the storing of working memory. The fourth verbal-lesion model (VLMG) involved a global lesion including all previous verbal lesions.

Both the EPM and IM do not show any statistical difference in the behavioural indices compared to the control model. This result suggests that high functioning language component ($\lambda = 0.81$ in both cases) compensates low error sensitivity

	μ	ϕ	τ	λ
Control model	0.49	0.97	0.10	0.81
Extreme perseverative model (EPM)	***0.001***	0.97	0.10	0.81
Distracted model (DM)	0.49	0.97	***0.4***	0.81
Irrational model (IM)	0.49	***1.0***	0.10	0.81
First verbal-lesion model (VLM1)	0.49	0.97	0.10	***0.001*** *(only if r = 0)*
Second verbal-lesion model (VLM2)	0.49	0.97	0.10	***0.001*** *(only if r = 1)*
Third verbal-lesion model (VLM3)	0.49	0.97	0.10	***0.001****(r not considered)*
Global verbal-lesion model (VLMG)	0.49	0.97	0.10	***0.001***

Table 3.10: Parameters of the lesioned models obtained by altering the parameters of the control model that fits the human control group (data reported in Baldo et al., 2005). The first three models involve lesions of main processes of the model, while the last four models involve four different lesions of the language component. Values in ***bold Italics*** represent the parameters that were altered to produce the lesioned models.

(EPM has $\mu = 0.001$) and high forgetting (IM has $\phi = 1.0$), highlighting the vicarious role that language can play. Conversely, the DM shows a statistically significant difference for each index ($p < 0.001$ for each index). In particular it shows a lower CC (0.88 ± 0.9 vs 6.0 ± 0), a higher TE (47.65 ± 6.39 vs 12.35 ± 4.14), formed by 40% of PE and 60% of NPE, and a higher FMS (2.71 ± 1.7 vs 0.59 ± 0.69). This is coherent with the theoretical interpretation of the τ parameter representing distractibility since high distractability cannot be compensated for by language.

The VLM1 (impairment of the language contribution in case of negative feedback) showed a statistical significant difference with each index of the control model, with the exception of CC and FMS. In particular, the model showed a higher TE (20.41 ± 8.35 vs 12.35 ± 4.14, $p < 0.01$), composed by 48% of PE and 52% of NPE. Overall, this lesioned model did not show a great global impairment, as shown by the high CC and a slightly higher number of total errors. The reason is that in this model the impairment only corrupts the language contribution in case of negative feedback and this is compensated by an intact high error-sensitivity coefficient ($\mu = 0.49$).

The VLM2 (impairment of the language contribution in case of positive feedback) showed worse performances compared to both the INVM and control models. In particular, it showed a very low CC (0.06 ± 0.24 vs 6.0 ± 0, $p < 0.001$) and a

very high TE (54.71 ± 4.08 vs 12.35 ± 4.14, $p < 0.001$) formed by 35% of PE and 65% of NPE. With the exception of FMS, that is not statistically different from the one of the control model, the errors profile of this model are similar to those of the DS model ($\tau = 0.4$), showing a high number of errors and in particular an imbalance toward NPE, thus suggesting an attention impairment. This suggests that language plays an 'attentional focus' function that in case of positive feedback increases the probability of focusing on the specific correct rule discovered and stored in memory.

The model VLM3 (impairment of the language contribution to the storing function based on the phonological-loop) did not show any statistical difference in any behavioural index compared to the control model. This suggests that the simple storing function of inner speech has not a relevant role in this task.

The VLMG (global impairment of the language contribution) exhibits the worst indexes with respect to all models. In particular, the model has a very low CC (0.12 ± 0.32 vs 6.0 ± 0, $p < 0.001$) and the highest TE (60.82 ± 4.66 vs 12.35 ± 4.14, $p < 0.001$), formed by 35 % of PE and 65% of NPE, and a higher FMS (1.24 ± 0.94 vs 0.59 ± 0.69, $p < 0.05$). The error profile of this model is similar to the one of the DM but shows worse indices (with the exception of FMS). This result suggests that a global impairment of the language system causes a severe deterioration of the model flexible goal-directed behaviour.

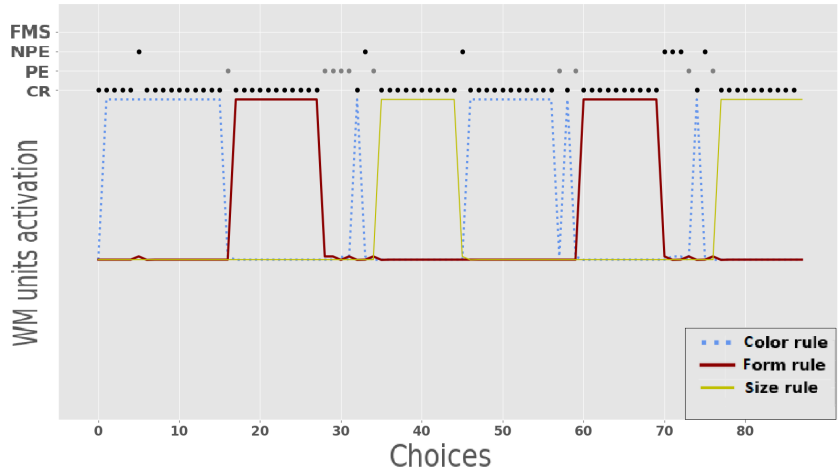
Analysis of internal functioning of the model

Here I show the internal functioning of the control group (Figure 3.17) and the three models each with a specific lesion of the inner-speech component (VLM1, VLM2, VLMG; Figure 3.16).

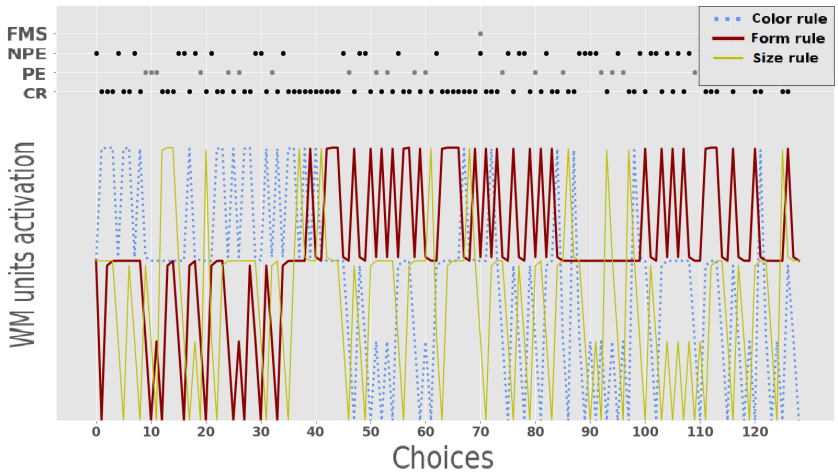
I do not consider the model with an impaired phonological loop function (VLM3) because it did not show any statistically relevant difference with the control model.

I also show a plot related to a model without the language system and fitting the data of the human control group. This model shows what happens if one assumes

Verbal-lesion model 1 (lesion of language negative feedback processing)



Verbal-lesion model 2 (lesion of language positive feedback processing)



Global verbal-lesion model (lesion of all language functions)

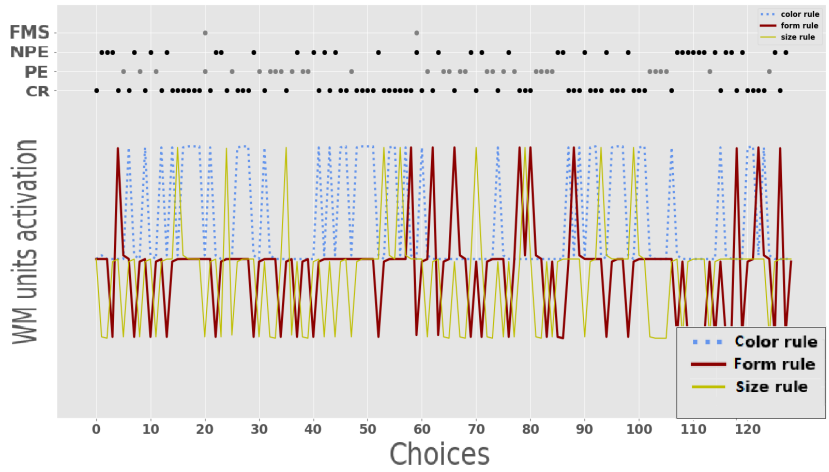


Figure 3.16: Internal functioning of the three models with lesions affecting different functions of the inner-speech component (Verbal-lesion model 1, Verbal-lesion model 2, Global verbal-lesion model). Each line in the graphs shows the activation of a working-memory unit representing a tendency to choose a specific sorting rule between the three possible rules. The dots at the top of graphs indicate single instances of correct responses (CR) or errors (PE, NPE, FMS).

that during development the absence of the inner-speech component would be substituted by vicarious cognitive processes still supporting flexible behaviour. Note that the different models might also capture individual differences in the use of inner-speech as a support for high-level cognitive processes, as highlighted in Baldo et al. (2005). The two control models have a similar good performance (they both solve the task in 80 rounds) but they exhibit a partially different internal functioning that highlights the role that language can play in the solution of the task (Figure 3.17).

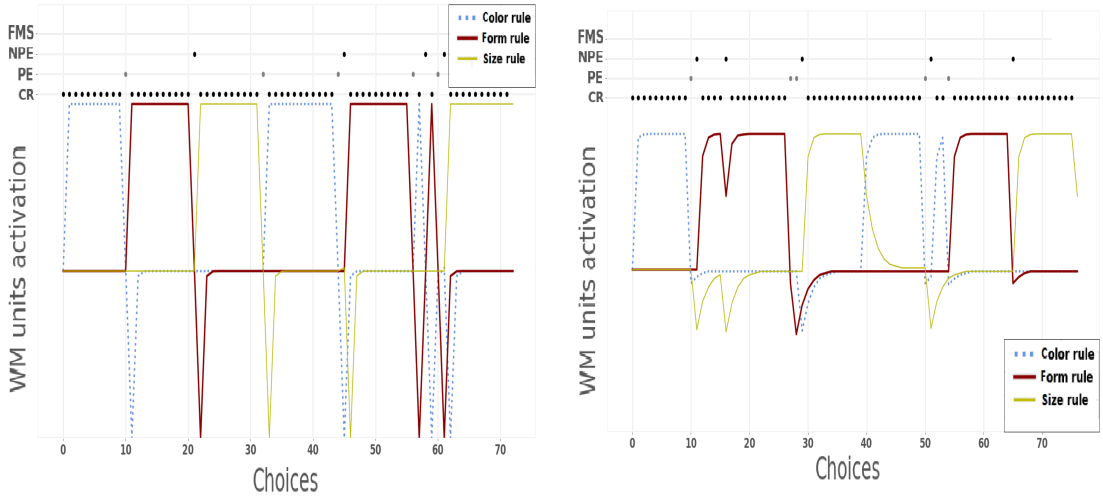


Figure 3.17: Internal functioning of two control models. Left: model with language. Right: model without language. Each line in the graphs shows the activation of a working-memory unit representing a tendency to choose a specific sorting rule between the three possible rules. The dots at the top of graphs indicate single instances of correct responses (CR) or errors (PE, NPE, FMS).

Notwithstanding the similar behaviour, the qualitative comparison of the activation of the working-memory units encoding the different decision rules shows that they are more 'disentangled' in the model with the language component. In particular, in the model with language the units have a more polarised activation and sharper activation changes. This supports a higher cognitive flexibility, in particular rapid decreases of activation after an error, and a high focused capacity, in particular rapid increases of activation after a positive feedback.

The VLM1 (language impairment in case of negative feedback) has an internal functioning similar to the one of the control model, but also some differences.

In particular, the high information forgetting speed ($\phi = 0.97$, shared with the control model) causes a fast decay of the activation to 0.5 (baseline activation of units) while the absence of a linguistic contribution in case of negative feedback prevents strong decrements of the activation of units. The effect of these two specific processes causes a minor difference between the activation of the units (in particular an higher inferior boundary) thus producing more PE (e.g., see choice interval 28-31) and NPE (e.g., see choice interval 70-75).

The VLM2 (language impairment in case of positive feedback) shows an erratic and inefficient internal functioning. In particular, it shows a lower superior bound of activation and sudden extreme activation changes that do not allow the completion of the test and produce several PE and NPE. Paradoxically, the erratic behaviour causes also random completions of categories (e.g., see choice interval 35-44). This result shows that the language contribution after a positive feedback supports the focus on a specific rule for prolonged times.

The VLMG (all language impairments) showed an internal behaviour similar to an average behaviour of the previous two models. In particular, it exhibits a minor range of activation (inferior and superior bounds) and erratic changes of activities that prevent the completion of any category. Interestingly, this plot is qualitatively more similar to the one of VLM2 (positive feedback and focusing impairment) than to the one of VLM1 (negative feedback impairment), thus corroborating the previously discussed quantitative data indicating that the language contribution to focusing is more important than its contribution to processing feedback errors.

3.2.4 Discussion

The model proposed here highlights the specific mechanisms through which inner speech might enhance the internal manipulation of representations involved in goal-directed cognitive processes and executive functions. In particular, it accounts for the cognitive flexibility as measured in the Wisconsin Card Sorting Test (WCST).

Theory-driven statistical analyses, focusing on versions of the model having parameters involving a negligible role of language, highlighted the similarity of the functions played by feedback-based working-memory update processes by language capable of supporting the *attentional focusing on successful goals* (behavioural rules). In addition, the comparison between the control model and versions of the model whose language *storing function* was lesioned did not show significant statistical differences, thus suggesting that the support of language for this function has a negligible role in the target experimental test. These results corroborate the importance of working-memory in flexible cognition as measured in the WCST, highlighting its role of 'executive function' rather then of mere information storage (Barceló & Knight, 2002). The results presented here thus suggest that inner speech can play a key role in enhancing the capacity of manipulating the high-order representations of the model (i.e. the states of working memory) and thus to improve the effectiveness of goal-directed behaviour.

In my work I also analysed the internal functioning and interaction of the system components. This analysis involves the control model with and without language and the language-lesioned versions of the model. Results indicate that the role of inner speech in to enhance the executive-function role of working memory is based on its capacity to strengthen the activation differences between neural units which represent alternative possible goals (the behavioural rules to follow in the solution of the WCST). Once so differentiated, the internal representation carrying relevant information are more robust to lesions, distractions, and internal/external sources of noise. This 'disentanglement' function of language also manifested in a previous abstract non-embodied computational model (Mirolli & Parisi, 2006) further discussed in section 3.2.4. Interestingly, artificial intelligence has recently started to highlight and study the importance for neural-network architecture be able to 'disentangle' internal representations to enhance the signal/noise ratio for downstream processing components (Goodfellow et al., 2017; Zhang & Zhu, 2018).

Overall, these results corroborated a *super-ordinate role* of inner speech that involves both executive functions and perceptual embodied processes. In this perspective,

inner speech represents a boosting internal cognitive 'tool', as highlighted by the psychological literature discussed in Section 2.3, executing a second-order manipulation of high-order representations.

Comparison with other models of language-cognition interaction

Few scientific studies focus on the interaction between language and cognitive processes adopting a theoretical and computational approach. Here I review computational models that contribute to investigate this important topic.

A first biologically grounded model simulates the brain networks supporting the interaction between attention and language (Garagnani et al., 2008). This model has high biological fidelity, simulating the neurophysiologically and anatomically grounded networks supporting the perception and production of speech.

The model uses winner-take-all neural mechanisms to reproduce bottom-up attentional selection of words and non-words at different brain levels. However, the model does not investigate how goal-based top-down processes might affect internal representations. Another computionally more sophisticated model (Garagnani & Pulvermüller, 2013), expanding the previous model, simulates the brain neural networks supporting decision-making processes for action and speech. The model is used to explain the spontaneous emergence of intentional speech acts in the brain. The model has the same limitations of the previous model and in addition investigates the influence of high-order processes on language but not the influence of language on cognition as done here.

More abstract models investigate the interaction between language and cognition in particular focusing on the supporting role played by language for categorisation processes. A first model (Cangelosi et al., 2000) links symbolic processing (words) to neural distributed representations and implements deep neural-network architectures involving sensory-motor learning and symbolic learning. The model investigates the top-down effect of symbolic computations on neural-network representations, suggesting that language can represent a 'symbolic theft' tool

to improve categorisation. This model is not validated with empirical data as here and addresses a different investigation problem with respect to the role of language for the internal manipulation of representations.

Three further models show mechanisms that use a 'linguistic labels' to influence the neural networks computations. The first model (Lupyan, 2005) is based on an auto-encoder neural-network. The model is used to show how the injection of linguistic labels into the intermediate layers of the model can enhance its classification capabilities. The model has an abstract architecture that cannot be mapped onto specific cognitive processes and is not validated with empirical data, but it nevertheless proposes an interesting mechanism for which language can manipulate the internal representations of auto-encoder neural networks. Another computational model (Mirolli & Parisi, 2006) highlights a possible role of inner speech for cognition. The model in particular includes two simple neural networks that model a sensory-motor loop, learning to categorise objects, and a phonological loop, learning to repeat words. The two networks interact at the level of their intermediate layers. Although the simple architecture does not capture specific cognitive processes and the model is not validated with empirical data, the results show that self-directed language can enhance the disentanglement of internal representations for different categories of objects, a phenomenon also emerged here. A last model (Caligiore et al., 2010) solves a sensory-motor classification tasks and shows how language can be used with vicarious functions with respect to visual inputs and to activate goals allowing to flexibly respond to stimuli. The model is qualitatively validated with empirical data but it does not investigate how language might influence the manipulation of internal representations.

Overall, compared to all aforementioned models the model presented here shows an an architecture directly capturing key high-level cognition processes. Moreover, it allows the study of how language can act as a cognitive tool supporting the manipulation of internal representations enhancing the interaction with the external environment. In so doing, it focuses on a superordinate role of language that can potentially explain its influence on many different domain-specific tasks (e.g.

object categorisation or cognitive flexibility). Due to these differences the model represents a more recent and complex model of self-directed language/cognition interaction.

3.2.5 Conclusions

This study corroborates and extends the three-component hypothesis, suggesting that a self-directed form of manipulation of high-order representations (e.g. inner speech) can participate to the expression of a flexible cognition and behaviour. In particular, the model is validated reproducing human behavioural data during the performance of the Wisconsin Card Sorting Test, designed to study cognitive flexibility, both in a standard condition and in a condition involving verbal shadowing. Furthermore, the analysis show how language can play multiple functions to support an high-order representations manipulation, such as the processing of external feedback and working memory. In particular, the inner speech ameliorates attention engagement and disengagement with respect to specific goal representations after a feedback is received, and in general augments the disentanglement of goal representations.

3.3 Model 3. Motivated categorical perception: a precursor of internal manipulation

Here I introduce the third computational study that corroborates the motivated categorical perception theory, focusing on the representation learning processes at the basis of the acquisition of suitable perceptual representations. In particular, this section introduces the task, the computational components of model and the obtained results. At last I propose a discussion and conclusions about this study.

3.3.1 Task and experimental conditions

Overall, the task I used to test the model is inspired by category learning tasks, requiring the production of a response on the basis of specific visual features of stimuli such as colour, shape, and size (for an extended analysis of these tasks see Ashby & Maddox, 2005, 2011). In particular, I focused on a sub-class of these tasks in which a classification rule is fixed and the participant has to execute a motor action on the basis of the features of a card (Hanania & Smith, 2010). Note that despite the task is inspired to experimental protocols, the same learning processes I emulate could support the ecological development of infants categorical perception (Clifford et al., 2009; Galle & McMurray, 2014).

At an operational level, the experimental protocol is composed of a 'pre-task section' and a 'task performance section' (Figure 3.18A). In the first the environment chooses a specific sorting rule (i.e. colour, shape, or size) and creates a set of 'ideal vectors'. These vectors correspond to the output vectors that the model should produce in correspondence to a specific input and a specific sorting rule. In this way, in each trial a visual input is provided to the model and the environment computes a feedback (reward) on the basis of the distance between the model response and the ideal response (see section 3.3.3 for further details of this calculus). For example, in case the environment chooses 'colour' as a sorting rule, all inputs with a specific colour (i.e. red, green, blue, or yellow) will be associated with one of four ideal vectors. The second section of protocol is composed of many trials in which the model interacts with a virtual environment trough four phases (Figure 3.18A, on top). First, the environment provides a single visual input to the model, that processes it (phase 1). The visual input is extracted from a set of 2D input images of geometrical shapes varying in colour, shape, and size, produced from four example images (Figure 3.18B). Second, the model produces an output (distributed binary vector) on the basis of the processed visual input (phase 2). Third, the environment returns a score index that suggests the correctness of the model response with respect to the ideal one (phase 3). Fourth,

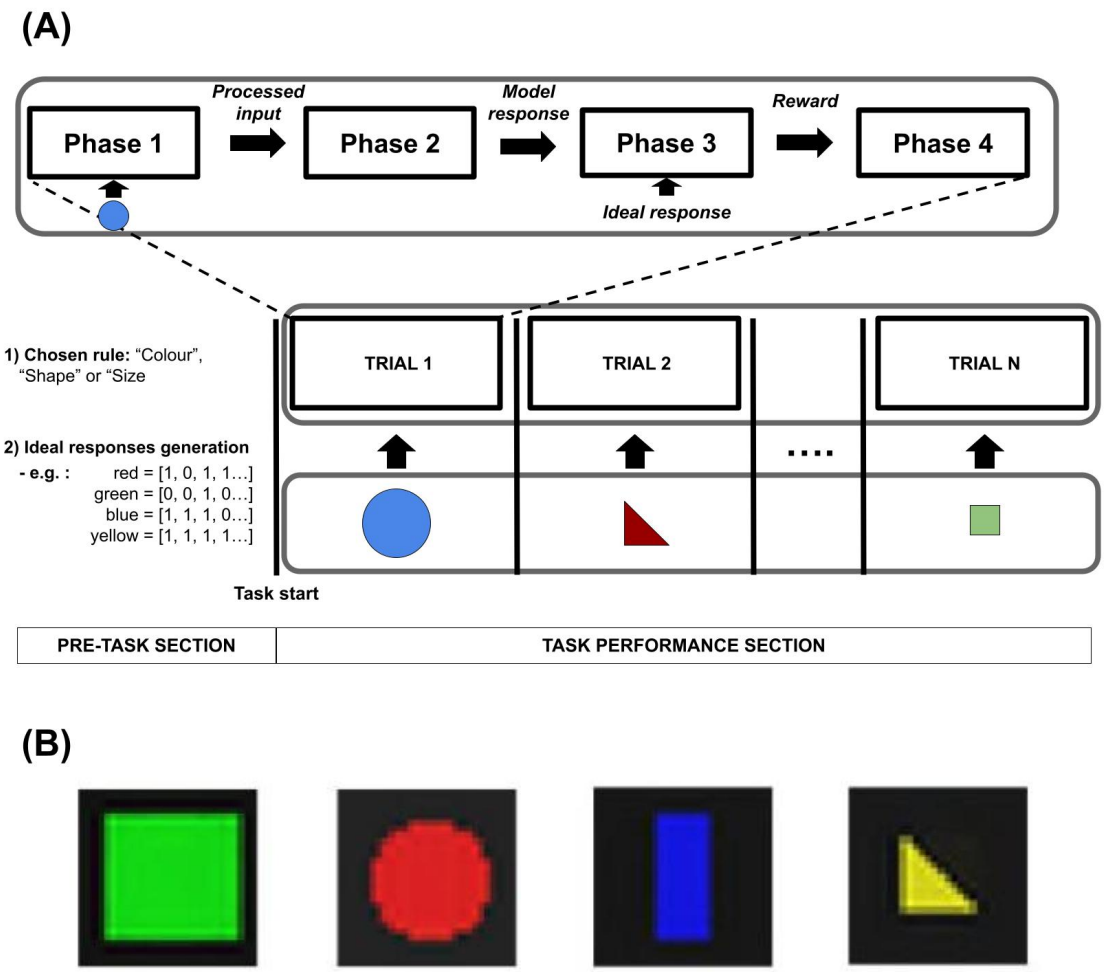


Figure 3.18: (A): Graphical representation of the task protocol. The row below shows the examples of inputs that the environment provides to the model (visual input). The middle row shows the trials sequence. Note that a first step occurs before the trials start and involves the setting of task conditions (i.e. choice of the sorting rule and creation of the ideal responses). The top row offers a 'zoom in' into a specific trial, showing the phases that occur during the model-environment interactions. (B): Examples of the 64 geometrical shapes (circles, squares, rectangles, triangles) used to produce the images. Each image encompasses a different attribute out of the four attributes of each of the three categories colour, shape, and size.

the model computes the reward returned from the environment and adapts its internal components (phase 4). Each trial is repeated for a fixed number of times in the same order without any change of the starting conditions, i.e. the sorting rule and hence the ideal responses.

3.3.2 Neuro-inspired underpinnings of the model: key components and dynamics

As in the first and second studies, the model abstracts the fine-grain biological details (e.g. neuronal micro-circuitry or bio-grounded plasticity). However, the interactions between the macro-systems underpinning the learning processes (e.g. motivational and perceptual systems interactions) are bio-plausible (e.g. localistic learning rule and distributed representations coding; Illing et al., 2019). This level of details is suitable for investigating the computational mechanisms that support the human learning processes underlying categorical perception.

Figure 3.19 shows the whole model architecture and the information flows between its components, also reporting the brain structures from which the components are inspired. Despite the model shows some simplifications, it proposes a system-level architecture that represents a promising approach in the computational modelling field (Eliasmith et al., 2012). The functional neural underpinning of the model components are now explained in-depth, while implementations details are reported in section 3.3.3.

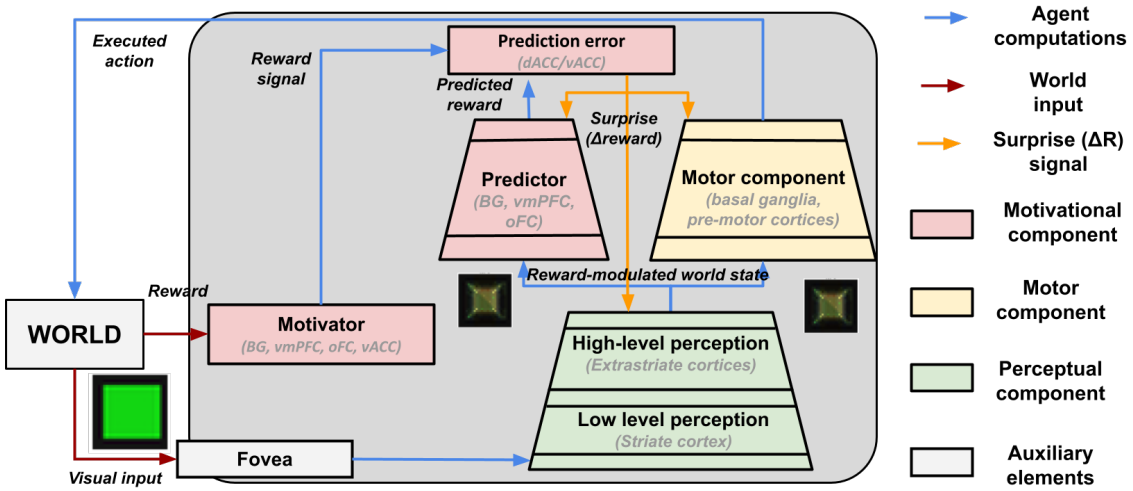


Figure 3.19: Schema of the model components and functions, the flows of information between the components, and the learning signals.

Perceptual component This component is based on a neural network that receives visual inputs and performs information abstraction, mimicking the brain visual system. In particular, the component emulates a hierarchical information processing (Felleman & Van Essen, 1991; Baldassarre et al., 2013a) from the low-level retinotopic features in striate cortex to the high-level features (e.g. colour, shape, size) in extrastriate cortices (DeYoe et al., 1996; Konen & Kastner, 2008).

Differently from the biologically implausible gradient-descent methods, the network learns through a bio-plausible mechanism (Illing et al., 2019). In particular, the learning rules update each connection weight (synapse) on the basis of locally available information related to the pre-synaptic and post-synaptic units. The distributional coding of representations is another biologically plausible feature of the model. Indeed, information on each content (e.g., a percept) is encoded by many units of the layer, and each unit takes part in the representations of different contents. This encoding is more bio-plausible than localistic representations ('grandmother-cells'; McClelland & the PDPResearchGroup, 1986; Quiroga et al., 2008). Finally, the differences in learning processes of the model layers represent a further bio-plausible feature. In particular, the top layer of this component, emulating extrastriate cortices, is trained through a mechanism that integrates associative and reward-based RL (Figure 2.5B). Instead, the bottom layer of the component, which mimics early visual cortices, is trained before the task execution reflecting an early development (Siu & Murphy, 2018). Critical for the motivated categorical perception hypothesis, these features capture the essence of the different weights that reward signals (e.g. dopamine-based inputs) have onto extra-striate and striate cortices (Williams & Goldman-Rakic, 1993; Jacob & Nienborg, 2018; Impieri et al., 2019; Niu et al., 2020; Froudist-Walsh et al., 2020).

Motor component This component is supported by a neural network that, on the basis of the perceptual component activation, produces an 'action' affecting the world. The network is trained through a trial-and-error learning algorithm using a reward signal, mimicking the interactions of basal ganglia with motor

cortices during the learning of actions (Kim et al., 2017; Seger, 2008).

Motivational component This component is formed by three sub-modules that emulate the motivational functions supported by different brain sub-systems.

First, a *motivator* sub-module produces a reward signal on the basis of the action outcome. Here the outcome is received from the environment and informs the system on the 'correctness' of the performed action (see below). This action-outcome might correspond to an 'extrinsic reward' (e.g. food or other rewarding resources) and is suitably processed by the system sensors and motivator component to produce a reward signal. Alternatively, the reward signal might be produced by intrinsic motivation processes (Baldassarre & Mirolli, 2013; Baldassarre, 2011) related to the novelty or surprise of the experienced stimuli (Barto et al., 2013) or to the goal-directed acquisition of competence (White, 1959; Santucci et al., 2016). In the brain, sub-cortical and ventral cortical structures support extrinsic rewards (Panksepp, 1998; Mirolli et al., 2010) while other sub-cortical and dorsal cortical structures support the computation of intrinsic reward signals (Lisman & Grace, 2005; Ribas-Fernandes et al., 2011; Baldassarre, 2011).

Second, a *predictor* sub-module, based on a multi-layer neural network, uses the representations of the top layer of the perceptual component to predict the future rewards. This module functionally mimics the brain basal-ganglia striosomes (Houk et al., 1995).

Last, a *prediction error* sub-module integrates the obtained and predicted rewards and produces a learning signal ('surprise'). This signal influences the learning of the predictor, of the motor component and, most importantly, of the perceptual component. In the brain, this signal is represented by the phasic dopamine bursts reaching various target areas (Schultz, 2002), and it has been modelled by the actor-critic RL architecture (Barto, 1995).

3.3.3 Computational details of the model

The architecture (Figure 3.20) is formed by a generative model integrated into an actor-critic architecture (Sutton et al., 1998), both modified to study the role of unsupervised ad reinforcement learning supporting the emergence of categorical perception. Moreover, auxiliary computational elements support the interaction between the model and an abstract task protocol (e.g. the world feedback).

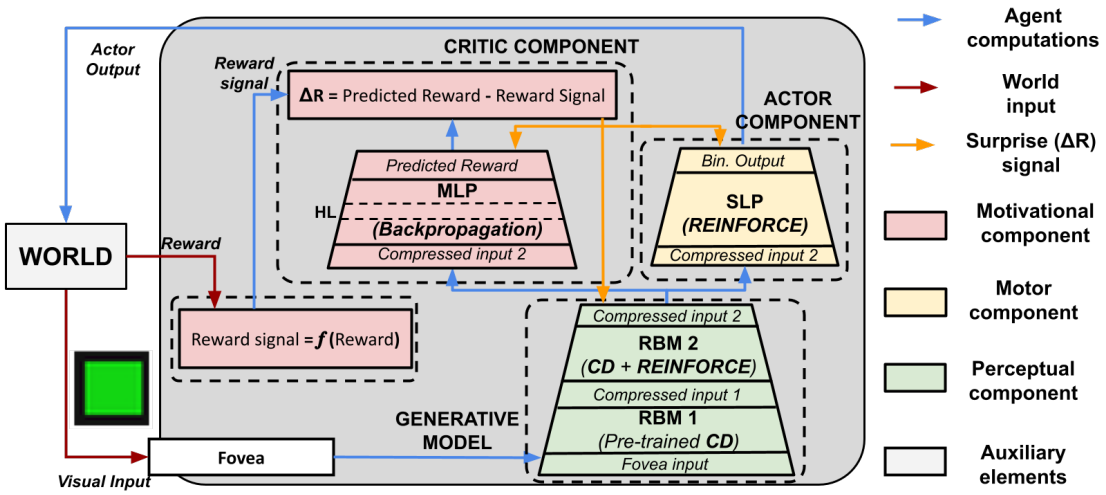


Figure 3.20: A computational schema of the model components and their training algorithms, the flows of information between the components, and the learning signals. MLP: Multi-layer Perceptron. SLP: Single-layer Perceptron. HL: Hidden Layer. RBM: Restricted Boltzmann Machine. CD: Contrastive Divergence.

Perceptual component This component is a generative *Deep Belief Network* (DBN; Hinton et al., 2006; Le Roux & Bengio, 2008) composed of two stacked *Restricted Boltzmann Machines* (RBM; Hinton, 2012). Each RBM is composed of an input layer ('visible layer') and a second layer ('hidden layer') formed by Bernoulli-logistic

stochastic units where each unit j has an activation $h_j \in \{0,1\}$:

$$h_j = \begin{cases} 1 \text{ if } v \geqslant \sigma(p_j) \\ 0 \text{ if } v < \sigma(p_j) \end{cases} \tag{3.5}$$

$$\sigma(p_j) = \frac{1}{1 + e^{-p_j}}$$

$$p_j = \sum_i (w_{ji} \cdot v_i)$$

where $\sigma(x)$ is the sigmoid function, p_j is the activation potential of the unit h_j, v is a random number uniformly drawn from $(0,1)$ for each unit, and w_{ji} is the connection weight between the visible unit v_i and h_j. The RBM is capable of reconstructing the input by following an inverse activation from the hidden layer to the input layer.

The DBN consists of a stack of RBMs—two in the model—where each RBM receives as input the activation of the hidden latent layer of the previous RBM. The model is trained layer-wise, starting from the RBM which receives inputs from the environment and towards the inner layers. On this basis, the DBN executes an incremental dimensionality reduction of the input, as higher layers further compress the representations received from the lower/previous RBM (Hinton & Salakhutdinov, 2006). In the model, the first RBM directly receives the input images and it is trained to encode them 'offline' before the task. This training adopts the *Contrastive Divergence* (CD), an unsupervised-learning algorithm that computes each connection weight update Δw_{ij} on the basis of a bidirectional iterative process. In particular, the visible layer receives an external input and activates the hidden layer, that in turn re-activated the previous visible layer (the weights of an RBM are bidirectional). Then, this reactivated visible layer activates the hidden layer for the second time. This cycle, constituted by a direct and inverse spread of the input, can be repeated many times but it usually fixed to have two activations of both visible and hidden layers. The first activations of the visible layer and the hidden layer are usually labelled as 'data' activations, in that are directly caused by the the external data (input). Differently, the activations that

are not directly caused by the original input (first spread of the network) are usually labelled as 'model activations' or 'reconstructions'. The following formula describes the CD algorithm:

$$\Delta w_{ij} = \rangle(\langle v_i \cdot h_j \rangle_{data} - \langle v_i \cdot h_j \rangle_{model}) \tag{3.6}$$

where ϵ is the learning rate, $\langle v_i \cdot h_j \rangle_{data}$ is the product between the initial input (initial visible activation) and the consequent hidden activation, $\langle v_i \cdot h_j \rangle_{model}$ is the product between the reconstructed visible activation and a second activation of the hidden layer following it, averaged over all data points.

The second RBM of the model is trained 'online' during the task performance based on the novel algorithm proposed here. The algorithm integrates *Contrastive Divergence* (Eq. 3.6) with the *REINFORCE* algorithm described in the next session (Eq. 3.8) as follows:

$$\begin{aligned} \Delta w_{ij} = \lambda\ (\epsilon\ (\langle v_i \cdot h_j \rangle_{data} - \langle v_i \cdot h_j \rangle_{model})) + \\ (1-\lambda)\ (\alpha\ (r - \bar{r})(y_j - p_j)x_i) \end{aligned} \tag{3.7}$$

where λ is the contribution of Contrastive Divergence to the update of weights, and $(1-\lambda)$ the contribution of REINFORCE. Crucial for this work, λ mixes the contribution of UL and RL processes to the weight update, in particular a high value implies a dominance of UL whereas a low value implies a dominance of RL. In the simulations, I tested five values of the parameter: $\lambda \in \{1, 0.1, 0.01, 0.001, 0\}$.

Motor component This component is a single-layer perceptron trained with the RL algorithm REINFORCE (Williams, 1992). The input of the network is the activation of the last layer of the perceptual component. The network output layer is composed of Bernoulli-logistic units as for the perceptual component. The algorithm computes the update Δw_{ji} of each connection weight linking the input

unit i and the output unit j of the component as follows:

$$\Delta w_{ji} = {}_{,}(r - \bar{r})(y_j - ff(p_j))x_i \tag{3.8}$$

where α is the learning rate, r is the reward signal received from the motivator, $\bar{r}$ is the reward signal expected by the predictor, x_i is the input of the network (from the outer second hidden layer of the DBN), $\sigma(p_j)$ is the sigmoidal activation potential of the unit encoding its probability of firing, and y_j is the unit binary activation.

Motivational component This component implements the functions of the *critic* component of an *actor-critic* architecture (Sutton et al., 1998).

The *motivator* module computes the reward signal by scaling the reward perceived from the external environment into a standard value, the *reward signal* $r \in (0, 1)$:

$$r = f(Reward) \tag{3.9}$$

where $Reward$ is the reward perceived from the environment and $f(.)$ is a linear scaling function ensuring that the reward signal ranges between 0, corresponding to a wrong action, to 1, corresponding to an optimal action. This reward signal represents the pivotal guidance of the RL processes. As discussed in the previous sub-section, in other cases the motivator may involve further mechanisms, computing the reward signals on the basis of extrinsic and/or intrinsic motivation mechanisms.

The *predictor* module is a multi-layer perceptron composed of an input layer, an hidden layer, and an output layer. The input layer corresponds to the second hidden layer of the DBN while the output layer, composed of a single linear unit, corresponds to the expected reward signal $\bar{r}$ computed on the basis of the DBN activation. The perceptron is trained with a standard gradient descent method (McClelland & the PDPResearchGroup, 1986; Amari, 1993) using a learning rate α and the error e computed by the prediction-error component.

The *prediction error* module is a function that computes the reward prediction error (surprise) e as follows:

$$e = r - \bar{r} \tag{3.10}$$

where r is the reward signal from the motivator, and $\bar{r}$ is the expected reward signal produced by the evaluator. This error is used to train the predictor itself, the motor component, and the perceptual component.

Auxiliary elements The input dataset is formed by RGB images with a black background and a polygon at the centre (Figure 3.18B). The polygon is characterised by a unique combination of specific attributes chosen from three visual categories: colour, form, and size. There are four attributes for each category: red, green, blue, yellow (colour); square, circle, triangle, bar (shape); large, medium-large, medium-small, small (size). These attributes generate $4^3 = 64$ combinations forming the images used in the test.

The retina component is implemented as a $28 \times 28 \times 3$ matrix containing the RGB visual input. The matrix is unrolled into a vector of 2,352 elements that represents the input of the perceptual component.

The environment is implemented as a function that (1) chooses the correct sorting rule before the task performance and creates a set of ideal actions for each input, and (2) provides an image to the model at each trial. In every trial the model perceives and processes one input image (Figure 3.18A) and undergoes a cycle of the aforementioned learning processes based on the reward received from the environment after the action performance and Figure 2.5A). Here the environment computes the reward r' simply on the basis of the Euclidean distance between the model action and an 'optimal action':

$$Reward = \|\mathbf{y}^* - \mathbf{y}\|_1 \tag{3.11}$$

where $\mathbf{y}^*$ is the optimal action binary vector that the model should produce for the current input, $\mathbf{y}$ is the model binary action, and $\|.\|_1$ is the L1 norm of the vectors

difference. The optimal actions are four binary random vectors that the model should produce in correspondence to the items of the four input categories of the given task.

3.3.4 Results

I tested the model with different task conditions and model configurations. First I varied the sorting rule, hence the task shows three task conditions. For example, a specific task condition required sorting the cards by colour and another one by shape or by size. Note that the sorting rule is fixed before the task start and it does not change during the task performance. Second, I tested the model with five different levels of UL/RL contribution (λ parameter, see Section 3.3.3). This variation gave rise to five model conditions, labelled as follows: Level 0 (L0): no RL (i.e., only UL); Level 1 (L1): low RL; Level 2 (L2): moderate RL; Level 3 (L3): high RL; Level 4 (L4): extreme RL (no UL). Third, I tested the model with two further conditions, namely 10 and 50 units in the second DBN hidden layer. These conditions aim to test the impact of the computational resources (i.e. the number of suitable units for storing the input information) on the task performances.

I varied the parameters of these environmental and model conditions with a random grid search based on over 1000 simulations. The simulations were run in the *Neuroscience Gateway platform* (Sivagnanam et al., 2013).

The presentation of results is organised in three parts. The first part investigates the relationship between the specific UL/RL balances and the task performance. The second part investigates the relationship between the specific UL/RL balances and the nature of the perceptual representations acquired. Finally, the third part presents a graphical visualisation of the previous representations and an analysis on the amount of information (visual details) they stored.

Performances analysis

Figure 3.21 shows the training curves of the models, trained with different RL contributions in 15,000 epochs. The L0 models, using only UL, learn faster during the first 1,000 epochs but exhibits the worst final performance. Instead, the highest final performance is achieved by the L3 and L2 models where UL and RL are better balanced. Figure 3.22 shows the final performance of the models, namely

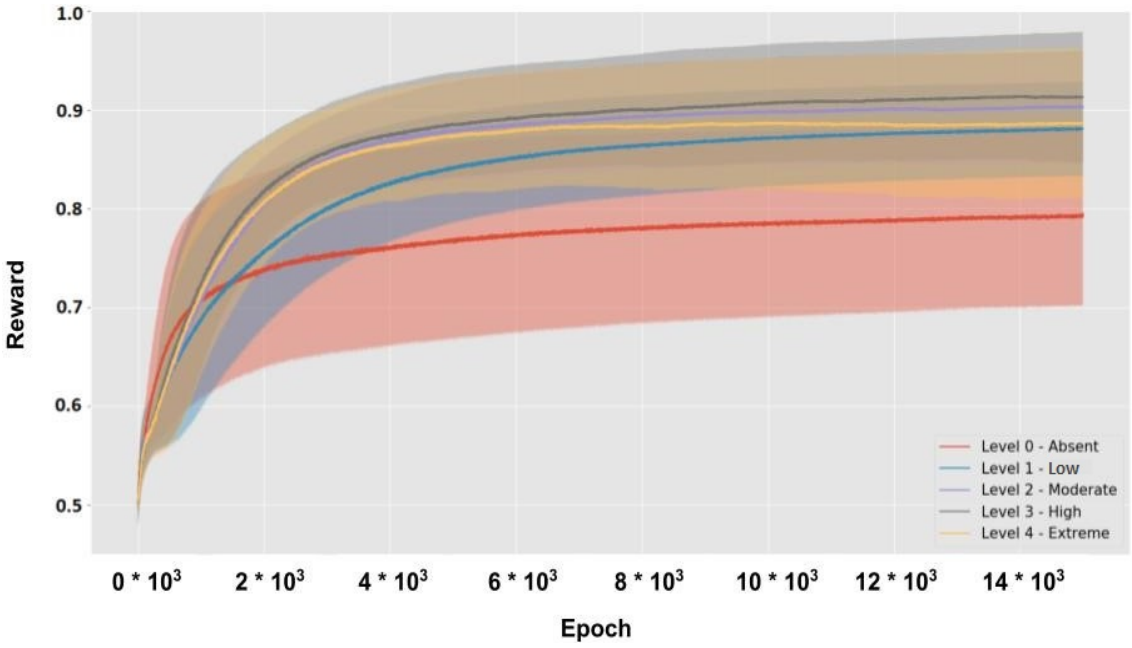


Figure 3.21: Reward per epoch of the five models involving different UL/RL levels, averaged over the models using a given level. Shaded areas represent the curves standard deviations.

the maximum reward they achieved.

A one-way ANOVA confirms the presence of a statistical difference between the final performance of the five groups ($F = 47.51$, $p < 0.001$). Post hoc tests (Table 3.11) confirm that the performances of models with an absent RL contribution (L0) are statistically different with respect to each of the other models (0.81 ± 0.08, all $p < 0.001$). The L3 models show a higher performance compared to the L0 models (0.92 ± 0.06 vs. 0.81 ± 0.08, $p < 0.001$), the L1 models (0.92 ± 0.06 vs. 0.89 ± 0.04, $p < 0.001$), and the L4 models (0.92 ± 0.06 vs. 0.90 ± 0.07, $p < 0.05$). The L2 and L3 models do not show a significant difference (0.92 ± 0.06 vs. 0.91 ± 0.05).

To further investigate the relationship between the performance of the models and

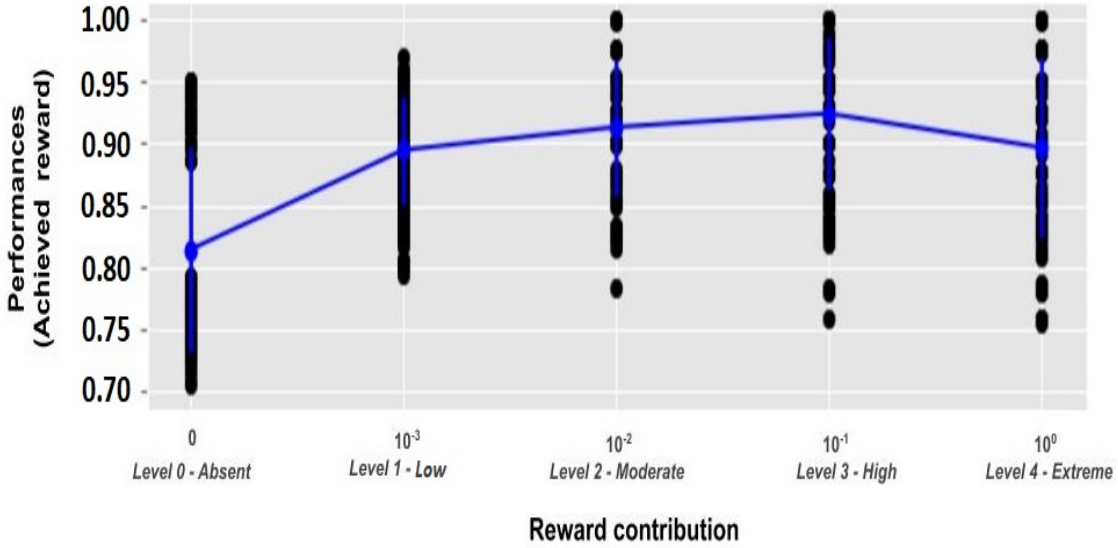


Figure 3.22: Performances (maximum reward obtained at the end of training) of models featuring different levels of RL contribution.

	Absent (L0)	**Low (L1)**	**Moderate (L2)**	**High (L3)**	**Extreme (L4)**
Absent (L0)	//	//	//	//	//
Low (L1)	$p < 0.001$	//	//	//	//
Moderate (L2)	$p < 0.001$	$p > 0.05$ (NS)	//	//	//
High (L3)	$p < 0.001$	$p < 0.001$	$p > 0.05$ (NS)	//	//
Extreme (L4)	$p < 0.001$	$p > 0.05$ (NS)	$p > 0.05$ (NS)	$p < 0.05$	//

Table 3.11: Post-hoc comparisons (t-test with Bonferroni correction) between the performance of models with different levels of RL contribution. 'NS' indicates 'non statistically significant'.

the different levels of RL contribution, I grouped the results of the simulations on the basis of the computational resources or the sorting rule (Table 3.12). Here I present a summary of the results.

Overall, increasing available computational resources tends to lower the amount of RL contribution needed to achieve the highest performance. Indeed, a one-way ANOVA shows a statistical difference between the models ($F = 3.85$, $p < 0.001$) and the post-hoc tests show that the L2 model leads to the best result (0.95 ± 0.05).

The table also highlights differences between the simulations using different sorting rules (colour, shape, size). The simulations with the *colour sorting rule* show flattened reward values with respect to the different RL contribution. In the case of low computational resources the model does not show statistically significant differences ($F = 0.88$, $p > 0.05$). A difference emerges in the case of

high computational resources ($F = 19.8$, $p < 0.001$) where the L2 models, having a balanced UL/RL mix, show the best final performance (0.98 ± 0.02).

The simulations with the *shape sorting rule* show statistical differences with both low computational resources ($F = 120.9$, $p < 0.001$) and high computational resources ($F = 20.4$, $p < 0.001$). In both cases, the models using a mixed level of UL and RL prevail: the extreme cases of the L0 models (only UL), and L4 models (only RL) have lower performances with respect to the L1, L2 and L3 models having a more balanced UL/RL mix.

Finally, the simulations with the *size sorting rule* show statistical differences with low computational resources ($F = 43.4$, $p < 0.001$) but not with 'high computational resources' ($F = 1.12$, $p > 0.05$). In the first case, the L0 models have the lowest performance.

	Absent	**Low**	**Moderate**	**High**	**Extreme**
Low Resources (Average)	0.81 ± 0.08	0.89 ± 0.04	0.91 ± 0.05	$\mathbf{0.92} \pm 0.06$	0.90 ± 0.07
Colour	0.92 ± 0.02	$\mathbf{0.92} \pm 0.02$	0.91 ± 0.04	0.91 ± 0.07	0.90 ± 0.08
Shape	0.75 ± 0.02	0.89 ± 0.04	0.94 ± 0.04	$\mathbf{0.95} \pm 0.04$	0.93 ± 0.06
Size	0.76 ± 0.02	0.88 ± 0.05	0.89 ± 0.06	$\mathbf{0.90} \pm 0.06$	0.86 ± 0.07
High Resources (Average)	0.92 ± 0.03	0.93 ± 0.04	$\mathbf{0.95} \pm 0.05$	0.93 ± 0.06	0.93 ± 0.05
Colour	0.94 ± 0.01	0.94 ± 0.01	$\mathbf{0.98} \pm 0.02$	0.95 ± 0.03	0.96 ± 0.02
Shape	0.93 ± 0.02	0.97 ± 0.02	$\mathbf{0.97} \pm 0.02$	0.96 ± 0.02	0.94 ± 0.02
Size	0.88 ± 0.02	0.88 ± 0.03	$\mathbf{0.90} \pm 0.05$	0.88 ± 0.07	0.88 ± 0.07

Table 3.12: Performance of models with different RL contributions in correspondence to two different amounts of computational resources (number of neurons in the second hidden layer of the DBN) and three different sorting rules (colour, shape, size). Labels with '(Average)' identify the average of the three conditions (colour, shape, size) in case of low or high resources. Values in bold highlight the highest value for each condition (along the rows).

Analysis of internal representations

To investigate the nature of the perceptual representations acquired by the models, I show the results of some example simulations with different sorting rules and different levels of the RL (other simulations lead to qualitatively similar results). Since I have adopted 'realistic inputs' (geometric figures), I have analysed the 'reconstructed representations' of the input layer rather than the hidden representations. I adopt this strategy to better interpret the acquired representations of the

original inputs, of which I can plot the original geometric images (see Figure 3.27). my sample tests on the representations in hidden layer show similar results .

To plot the representations I used a Principal Component Analysis (PCA), allowing a dimensionality reduction, and a K-means algorithm, supporting clustering. First, I extracted the first two principal components of the visible layer in correspondence to the original 64 input patterns. Second, The K-means algorithm was applied to the PCA results by setting $K = 4$, so that the algorithm grouped the representations into four classes, as the number of the actions.

Colour sorting category: reconstructed input

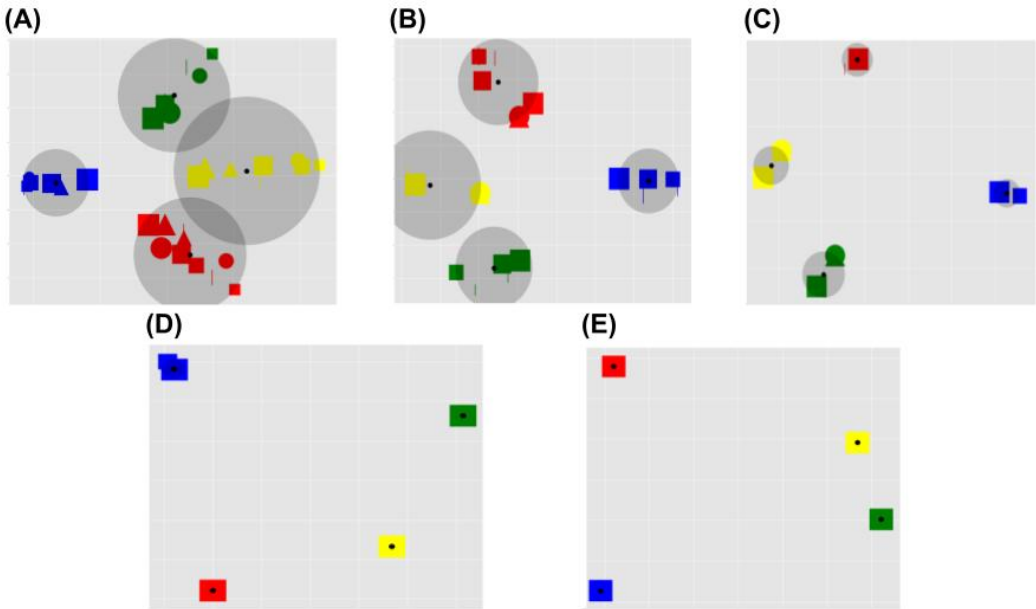


Figure 3.23: Principal components of the reconstructed image representations in the case of the colour sorting rule and in correspondence to different levels of RL (shown in different graphs). The dimensionality of the reconstructed image was reduced to two through a PCA (x-axis: first component; y-axis: second component). Within each graph, each reconstructed image is represented by a point marked by an icon that summarises the colour, shape, and size of the shape in the image (some icons are not visible as they overlap). The centroids of the four clusters found by the K-means algorithm are marked with a black dot, while the maximum distance of the points of the cluster from its centroid is shown by a grey circle. A: Level 0 (L0), absent RL (only UL); B: Level 1 (L1), low RL; C: Level 2 (L2), moderate RL; D: Level 3 (L3), high RL; E: Level 4 (L4), extreme RL (no UL).

The results (Figures 3.23-3.25) highlight that the RL contribution strongly affects the internal representations as revealed by the reconstructed inputs. Models with a medium (L2) and high (L3) level of RL show the emergence of task category-

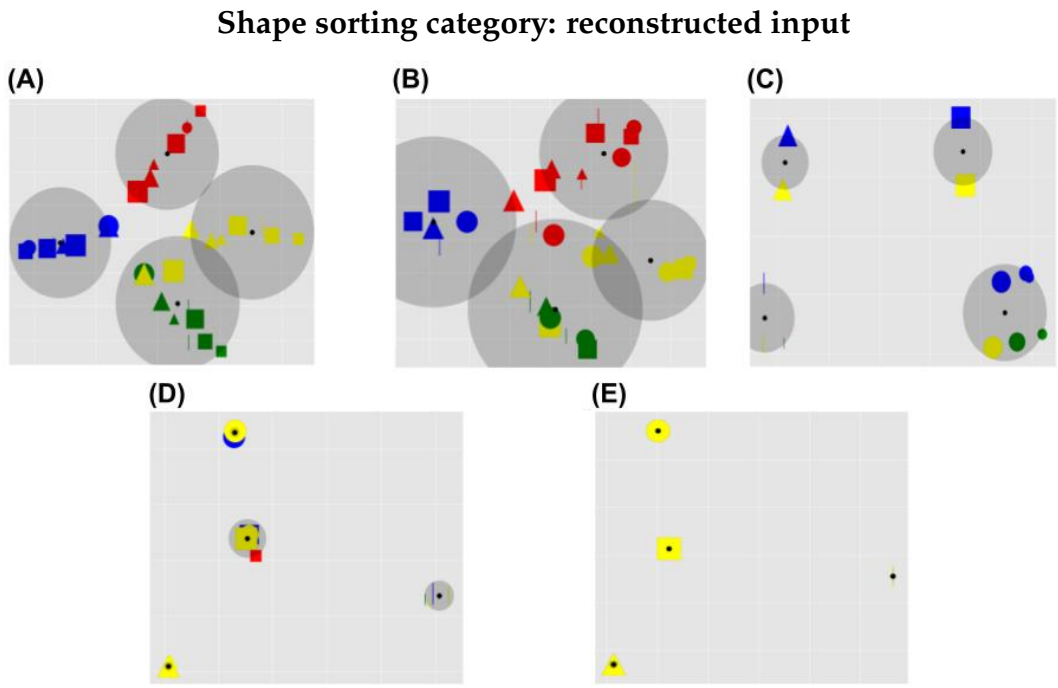


Figure 3.24: Principal components of the reconstructed image representations in the case of the shape sorting rule and in correspondence to different levels of RL. Note that, in case of overlap, the yellow inputs appear at the top and hide others due to technical factors (I plot the yellow inputs at the end). The plots are drawn as in Figure 3.23.

based clusters, whose radius progressively decreases as the weight of the RL increases. Conversely, the L0 and L1 models, with an absent or low RL, show a task-independent clustering effect on the basis of the input colours.

Figure 3.25-E shows that the model with an extreme RL incurred in a clustering error. In particular, in this condition the model should group the images into four clusters (as in the conditions of Figure 3.25-C,D) whereas it tends to use only three clusters and the fourth cluster on the right is almost empty.

Information stored by the model

To further investigate what type of information is stored by the perceptual representations, I show the results of two additional analyses. The first analysis examined the DBN reconstruction error while the second analysis qualitatively inspected the reconstructions of the input images.

Figure 3.26 shows the results of the first analysis and highlights the presence of a

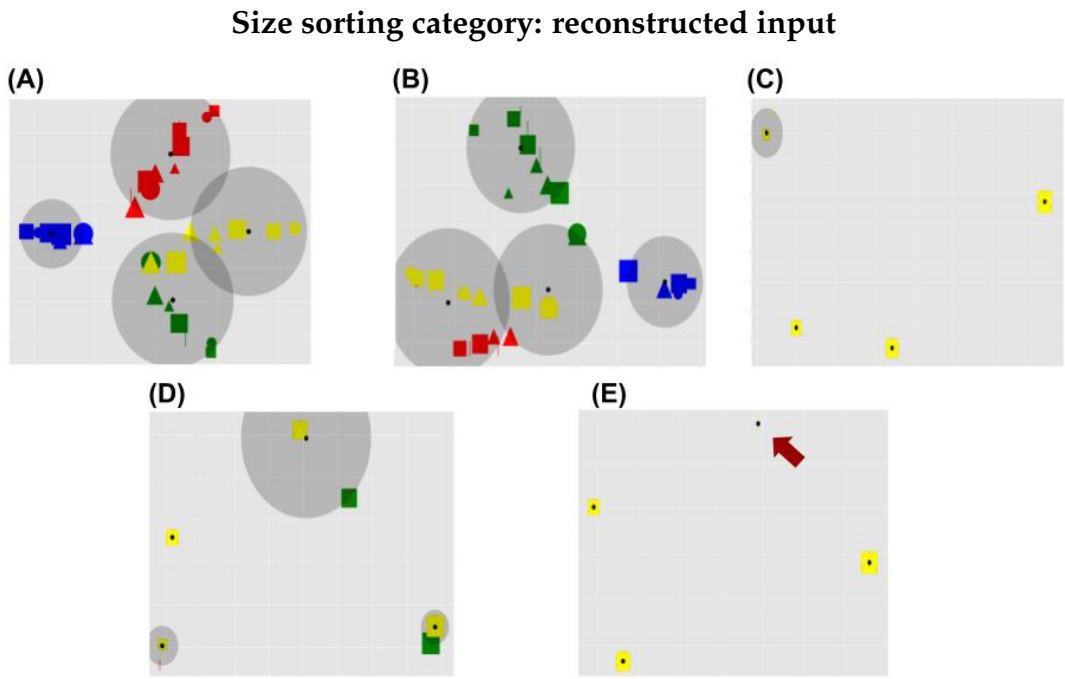


Figure 3.25: Principal components of the reconstructed image representations in the case of the size sorting rule and in correspondence to different levels of RL. Note that, in case of overlap, the yellow inputs appear at the top and hide others due to technical factors (I plot the yellow inputs at the end). The graphs are drawn as in Figure 3.23. The red arrow in graph E indicates the centroid of a cluster that contains only the small bars but not the other small shapes.

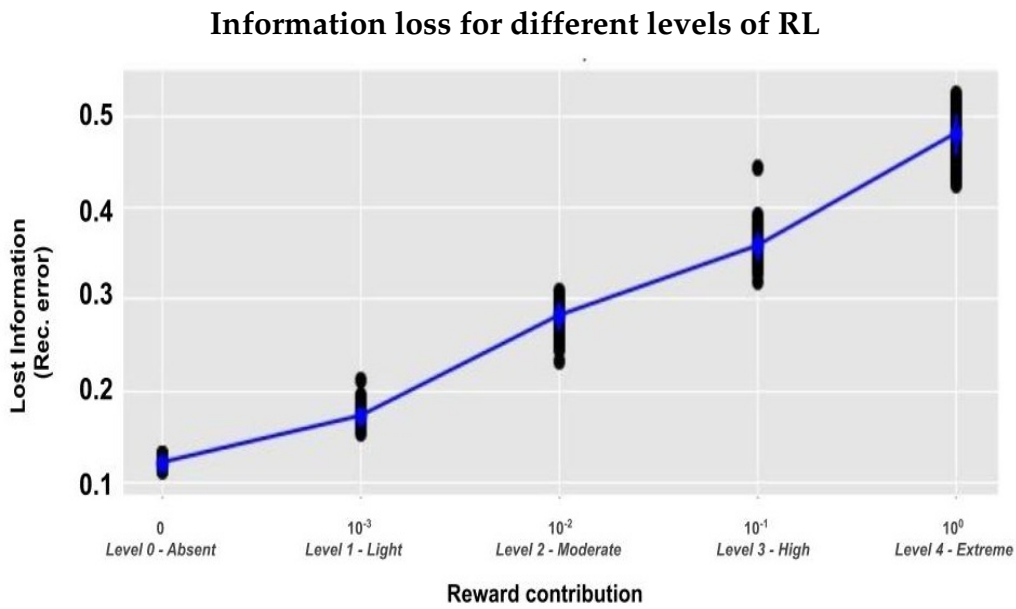


Figure 3.26: Information loss (reconstruction error at the end of the training) of models with different levels of RL.

strong positive linear relationship between the level of RL and the reconstruction error ($r = 0.68$, $p < 0.001$).

A one-way ANOVA confirmed the presence of a statistical difference between the five groups ($F > 100.0$, $p < 0.001$). These results indicate that an increasing RL contribution causes a progressive loss of information on the input images.

The qualitative inspection of the reconstructions shows the kind of information that the internal representations tend to retain, in particular if the system tends to store task-independent and/or task-related features. In this respect, Figure 3.27 highlights the emergence of a categorical perception, i.e. shapeless coloured blobs in case of colour sorting rule, colourless and sizeless prototypical shapes in case of shape sorting rule, and colourless blobs with different sizes in case of size sorting rule.

3.3.5 Discussion

Interpretation of the results

Here I discuss the results regarding the relationship between UL/RL contributions, behavioural performance and perceptual representations.

Unsupervised learning, reinforcement learning and categorisation performances. A main result of this work is that a suitable balanced mix of UL and RL leads the model to achieve the best performance in all tested conditions (Figure 3.22 and Table 3.11). Moreover, different UL/RL balances lead to different learning trends and behaviours of the models (Figure 3.21). For example, during the initial training phase the model with an absent reward contribution (L0) has some advantages, exhibiting the sharpest increasing learning curve with respect to the models with a higher RL (L2, L3 and L4).

A functional analysis of the models with a higher RL can explain this effect. These models initially produce a slow and highly variable exploratory behaviour, resulting in more early unstable perceptual representations. The early slowness and variability are caused by the key mechanisms of RL (Sutton & Barto, 2018), based on (1) an initial generation of noisy and stochastic representations, (2) a slow

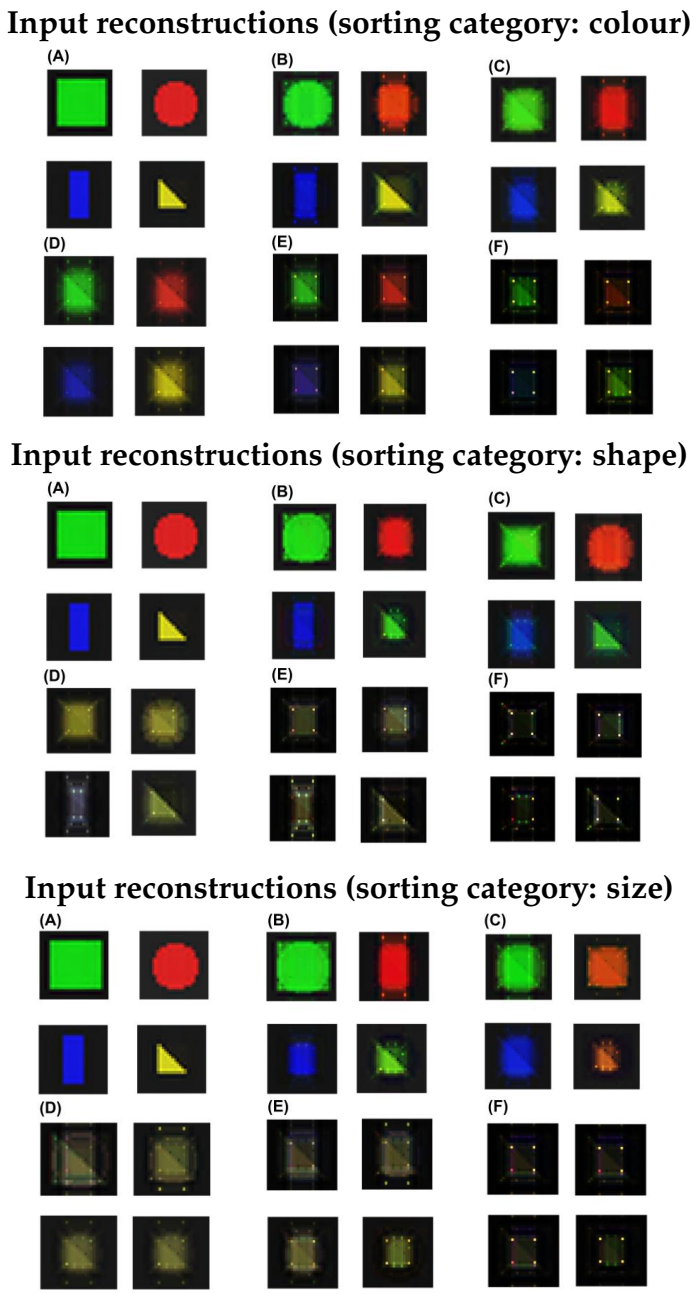


Figure 3.27: Image reconstructions with different sorting rules and different levels of RL. A: Original inputs; B: Level 0 (L0) - absent RL (only UL); C: Level 1 (L1) - low RL; D: Level 2 (L2) - moderate RL; E: Level 3 (L3) - high RL; F: Level 4 (L4) - extreme RL (only RL).

improvement in the prediction of the future reward (surprise) and (3) a representation learning based on both the stochastic generation and the surprise. Instead, the initial phases of the UL training can proceed regardless the slow learning to predict future reward (success of behaviour), and at the same time building suitable representations for the behaviour itself. However, with the advancement of training the conditions with absent RL (L0) and low RL (L1) achieve a lower performance than the more balanced conditions. This phenomenon occurs because in

the middle and last phases of training the other models (L2, L3 and L4) overcome the initial unstable phase, exploiting an higher task-directed bias (reward level) on the internal representations. Instead, the models with a low RL continue to encode both the task-independent and task-directed features without any specific bias. This unsupervised representation learning process is 'agnostic' with respect to the task performance and therefore causes a resources competition preventing the full exploitation of resources for task-directed computations.

At the opposite side of the spectrum, also models with an exclusive RL (extreme RL; L4) have computational limitations, resulting in sub-optimal performance (Figure 3.22 and Table 3.11). As also discussed above (Figure 3.21), the reconfiguration of the synaptic strengths is influenced by stochastic noisy activations and a slow reward prediction improvement. A consequence of these features is that these models show an inefficient initial representations learning, potentially incurring in local minima (e.g. Figure 3.25E).

These results are reproduced also by tests where I manipulated the computational resources that were available for the perceptual component (Table 3.12). These tests demonstrate that also with a higher amount of computational resources the best performance is achieved by the models having a balanced integration of UL and RL (moderate RL; L2). Interestingly, in case of higher resources the L2 model (moderate reward) shows the best performance while in case of low resources the L3 model (high reward) shows the best one. Despite this difference is small, there could be a functional explanation. Higher resource allow to encode more information helping to execute a correct categorisation. In particular, increasing the computational resource the UL mechanisms lead to store both more task-directed and task-irrelevant features, thus needing of a minor reward-based bias to tune the scarce resource toward task-directed feature (low resource condition). Nevertheless, 'storing all the information without a bias toward the useful one' remains an inefficient computational strategy due to a residual competition between task-relevant and task-irrelevant features. Hence, the L0 and L1 models show sub-optimal performance also in case of high resources. These results suggest

that also in case of high resources a trade-off between computational resource and task-directed bias leads to the best performance.

My results fit with the experimental evidences regarding the role of feedback signals in human adaptive behaviours. For example Soulières et al. (2007) suggest that in autistic people there could be an abnormal sensory processing. Corroborating experimental evidence (Frith, 2003), my results support the idea that in autism the feedback processing may be diminished, causing a certain level of autonomy of perceptual learning processes with respect to the task-dependent feedback. On the other hand, an abnormal reward sensitivity is considered one of the core factors bringing to clinical conditions as drug addiction or autism (Chelazzi et al., 2013; Mollick & Kober, 2020). Indeed, Seger & Miller (2010b) propose that autistic peoples could show an imbalance toward the reward-based plasticity, causing deficits in categorisation performances (e.g. low generalisation skills). my results corroborate this proposal, namely autistic people could show an excessive feedback-dependent sensory processing that causes sub-optimal performance with a potential loss of generalisation skills. Interestingly, the proposals of Soulières et al. (2007) (feedback insensitivity) and Seger & Miller (2010b) (excessive feedback-dependent sensory processing) seem in opposition. Future investigations could clarify this controversial evidence, however my results agree that (1) both sides of the imbalance can be detrimental to the categorisation task performance, and (2) the two imbalances could identify different categorisation profiles of the autism spectrum conditions.

Unsupervised learning, reinforcement learning, and categorical perception The main result is that different UL/RL interactions have a different impact on clustering process of internal perceptual representations, in turn leading to specific advantageous or disadvantageous effects on the task performance. In particular, in case of a balanced mix of the two learning processes (graphs 'C, D' of Figures 3.23, 3.24, and 3.25) a beneficial categorical perception effect emerges. Indeed, in this case the distances between inputs representations associated with a specific

response are reduced, while the distances between those associated with different response are expanded. This effect is made evident by the graphical reconstruction of original inputs (Figure 3.27). In case of a balance mixed of learning processes (e.g. graphs 'D' of the figure) the sensory system perceives the input as prototypes depending on the salient category (e.g., a coloured blob, when the task requires a colour-based categorisation, or a colourless shape prototype, when the task requires a shape-based categorisation).

These results corroborate the functional hypothesis proposed in the previous section. In particular, a balanced mix of unsupervised and reinforcement learning lead the internal representations to be clustered according to the task demands (categorical perception), thus improving action selection without losing salient information. Furthermore, the results are coherent with scientific evidences regarding the modulations of perceptual representations. For example, de Beeck et al. (2006) detect a training-dependent alteration of objects representation in human extrastriate cortices and Astafiev et al. (2004) detect a motor-related modulation of extrastriate cortices (in particular the exstrastriate body area). In addiction, Folstein et al. (2015) report that the solution of a category learning task causes the emergence of category-based representations. This phenomenon has been also shown in mice (Poort et al., 2015) and primates (Sigala & Logothetis, 2002; De Baene et al., 2008; Emadi & Esteky, 2014), thus indicating to have a key role along the evolution of mammal perceptual systems.

My model supports the investigation of imbalanced perceptual learning processes, leading to an absent or dysfunctional categorical perception. For example, in the case of absent or low RL (graphs 'A-B' of figures 3.23, 3.24, and 3.25) the unsupervised learning mechanisms lead to the acquisition of an high amount of visual features independently of their relevance for the task. This result is confirmed by the low reconstruction error obtained by these models (Figure 3.26), suggesting that they stores a higher amount of visual information. Moreover, the input reconstructions are very similar to the original inputs (graphs 'A, B' of Figure 3.27) confirming a very low loss of information. Interestingly, these models

show a certain level of clustering effect on the basis of the colour category due to the visual input coding. In particular, the UL mechanisms tend to extract the most preeminent statistical regularities and the colour coding is the most distinguishable feature of the inputs (e.g. many pixels code the colour of inputs while a few pixels differentiate the same-coloured shapes of blue circles and squares). Overall, these results agree with the functional hypothesis proposed in the previous section, for which task-independent perceptual representations can cause sub-optimal performance. Moreover, the emergence of a task-independent clustering effect could worsen the perceptual representation learning process.

At the opposite side of the spectrum, in the models with an extreme reward-dependent learning (graph 'E' of the figures 3.23, 3.24, and 3.25) the internal representations collapse to four specific ones, depending on the task demands. The highest reconstruction error (Figure 3.26) confirms an extreme information loss, sometime causing clustering errors (see graph 'E' of figure 3.25). Moreover, the input reconstructions (graphs 'F' of figure 3.27) offer a further evidence of the strong information loss. Indeed, the model can produce task-directed representations but they look less distinguishable with respect to those of the graphs 'D and E', sometime collapsing in a unique task-independent representation. Despite in this case the reward signal can support a task-dependent clustering effect, these models show a sub-optimal performance. This corroborates the idea that an extreme reward-based learning can give a general advantage to a perceptual system but it can also cause clustering errors. As detailed in the previous section, these disadvantages are caused by a slower and more variable learning mechanisms of RL. Moreover, in this case the UL/RL imbalance can cause a loss of useful information, potentially getting worse the generalisation skills.

These results could explain the proposals of Soulières et al. (2007) and Seger & Miller (2010a), suggesting that a weak top-down signal or extreme RL plasticity could affect categorisation and generalisation skill in autistic persons. Overall, the results I extracted from the 'extreme cases' could explain the altered computation in sensory cortices of autistic persons (Robertson et al., 2014; Humphreys et al.,

2008). Indeed, a recent review (Robertson & Baron-Cohen, 2017) proposes that an altered sensory computation in visual cortex is a key aspect to build better models of autism spectrum disorders. As explained in the next section, future investigations could clarify these experimental evidences.

Main contributions, clinical relevance and technological implications

Overall, my results propose many insights into the learning processes leading to categorical perception. First, a balanced contribution of unsupervised and reinforcement learning in high-order stages of a perceptual system leads to the best categorisation performance. This advantage is supported by a categorical perception effect, for which the perceptual system stores the visual information both on the basis of statistical regularities of inputs and task-dependent salience features. Second, the extreme cases of unsupervised and reinforcement representation learning lead to sub-optimal performances. In particular, exclusive unsupervised learning is inefficient due to an excessive autonomy of sensory computations with respect to the task demands. Instead, exclusive reinforcement learning causes a slow and variable sensory computation potentially leading to local minima of performance or clustering errors. These sub-optimal performance are caused by different alterations of perceptual representation learning. Indeed, in the first case the perceptual component stores too much information and hence shows a low task-directed CP effect. Conversely, in the second case the perceptual component acquires less distinguishable representations showing a maladaptive information loss.

The integration of my computational approach with specific experimental protocols, focusing on the feedback effect (Ashby & Maddox, 2011), and neuroimaging techniques, supporting the investigation of task-dependent sensory representations (de Beeck et al., 2006; Astafiev et al., 2004), could clarify the role of reward signals in healthy and clinical conditions of categorical perception. In particular, my model provides functional hypothesis and predictions about behavioural and imaging evidences. Indeed, my results suggest that the altered categorisation per-

formance in autism could be explained by an unstable categorical perception effect in extrastriate cortices, leading to sub-optimal generalisation skills and altered sensory computations (Humphreys et al., 2008; Robertson et al., 2014; Robertson & Baron-Cohen, 2017). For example, the 'extreme unsupervised learning' model, showing a maladaptive excessive autonomy between task demands and perceptual representation learning processes, corroborates a theoretical proposal explaining the altered categorisation process in autism (Soulières et al., 2007; Frith, 2003). However, the 'extreme reinforcement learning' model, reaching sub-optimal performances and potentially low generalisation skills, corroborates an alternative theoretical proposal for which autism could be supported by an extreme and inefficient reward-dependent representation learning (Seger & Miller, 2010a). Considering that autism spectrum condition shows many phenotypes in the social domain (e.g., iper-social and ipo-social profiles; Gao & Mack, 2021), my model reconciles the two opposing views suggesting that both the extreme UL/RL models corroborate the existence of different categorisation profiles in autism spectrum condition.

The computational principles and algorithms I used here can give a prompt to the machine learning and robotics fields. The field of reinforcement learning has a long tradition of studies that approaches the representation learning issue (Sutton et al., 1998; Caruana, 1997), also integrating UL and RL approaches (Jaderberg et al., 2016; Oord et al., 2018; Koutník et al., 2014). On the other hand, machine learning works propose many alternative architectures that aim to solve the representation learning issue. For example many studies adopt a variational auto-encoder (Kingma & Welling, 2019) also with practical applications (VAE; Wang & Gu, 2018; Sun et al., 2018). Moreover, recent approaches propose new variants of VAE such as the C-VAE (Sohn et al., 2015), approaching a multimodal representation learning framework, or the TD-VAE (Gregor et al., 2019), facing the sequential representational learning. Here I used a Deep Belief Network, composed of two Restricted Boltzmann Machines, that executes a representation learning and a dimensional reduction. I adopted this network due to specific computational and bio-inspired

features. First, VAEs are commonly implemented with Gaussian units while CD and REINFORCE are both natively implemented with Bernoulli units. This feature has allowed me to easily integrate the two algorithms in a single training equation of the DBN. Second, during the training the back-propagation influences each layer of the VAE while the DBN can be trained in a layer-wise way and each layer can be trained with a different algorithm. These features have allowed me to adopt different CD-REINFORCE balances along the DBN hierarchy, hence emulating the different impact of the reward at different stages of the brain sensory system. Third, CD and REINFORCE show a localistic learning rule that allows me to keep a certain level of bio-plausibility with respect to the back-propagation (Illing et al., 2019). Despite these features, ML approaches start to integrate many learning mechanisms to improve the efficiency of the representation learning process (Laskin et al., 2020; Chen et al., 2020). For example, Bengio et al. (2017) propose a first approach to train a VAE with a training function that integrates the back-propagation with a secondary object function, that potentially supports a reward signal. Future studies could explore the possibility to compare my DBN, trained with my novel algorithm, with a VAE trained with both back-prop and RL.

In addition to the previous studies, recent advances in deep learning (Mehrer et al., 2020; Bonnasse-Gahot & Nadal, 2020) and deep reinforcement learning (McInroe et al., 2021) are starting to elaborate indices to evaluate the task-related efficiency of representations, also investigating the issue of categorical perception in deep neural networks (Bonnasse-Gahot & Nadal, 2020). Taking inspiration from the different brain processes that support the representation learning in healthy and clinical human conditions, my approach can serve as a guide for these ML studies. For example, by analysing the categorisation deficits affecting humans in clinical conditions (e.g., autism) I could identify the latent causes that lead to generalisation limits in deep learning. On the side of robotics, some approaches (Böhmer et al., 2015; Parisi et al., 2017; Thomas et al., 2018) start to create learning functions integrating unsupervised learning and task-dependent reward functions, with the aim of better discriminating the visual features that provide the robot

with better control on the environment. Overall, my approach could prompt the construction of new robotic architectures, taking advantages of a balance between agnostic and task-directed perceptual processes (Posner, 2020; Cox et al., 2016).

Other computational models of categorical perception The computational literature concerning perceptual and learning processes is vast, involving many fields such as perceptual decision making, perceptual learning, category learning. Here I focus on categorical perception and I compare my model with other recent models that explicitly investigate this phenomenon (for a previous review of the categorical perception models see Damper & Harnad, 2000).

Models	**Computational features**			**Bio-plausible features**	
	Algorithms	**Learning mechanism**	**System-level approach**	**Architecture**	**Learning processes**
Beer (2003) Beer (2003)	Recurrent network	(Genetic algorithm)	✗	✗	✗
Spratling and Johnson (2006) Spratling & Johnson (2006)	Bio-constrained network	Unsupervised	✓ / ✗	✓	✓
Kröger et al. (2007) Kröger et al. (2007)	SOMs	Unsupervised	✓	✓	✓
Salminen et al. (2009)Salminen et al. (2009)	SOMs	Unsupervised	✗	✗	✓
Casey and Sowden (2012) Casey & Sowden (2012)	Bio-constrained network	Unsupervised	✓ / ✗	✓	✓
Tajima et al. (2016) Tajima et al. (2016)	RNN	(Bayesian inference)	✗	✗	✓ / ✗
Pérez-Gay et al. (2017) Pérez-Gay et al. (2017)	Autoencoder + MLP	Unsupervised, Supervised	✗	✗	✗
This model	Actor-Critic, Deep Belief Network, auxiliary components	Unsupervised, Unsupervised/Reinforcement	✓	✓	✓

Table 3.13: Overview of the main features of the computational models on categorical perception considered here. SOMs stands for self-organising maps;' MLP: Multi-layer perceptron. Entries in brackets under the respective column are not proper 'Learning mechanisms'. 'System-level approach' indicates whether the model emulates the computations of many brain structures beyond the perceptual component (e.g. subcortical structures). 'Bio-plausible features' indicates whether the model captures some aspects of the brain architecture (e.g., functioning of neurons and/or interactions of macro-systems) or learning processes (i.e., bio-plausible learning rules).

Beer (2003) proposes an evolutionary approach to model categorical perception effects. In this work an embodied agent, supported by a recurrent neural network and genetic algorithm, shows embodied loops with the world and evolves internal representations that support categorisation processes (embodied categorical perception). Despite the strong methodological differences with my proposal (e.g., the use of genetic algorithms), I share the interest in system-environment interactions and perceptual realism of the input leading to the emergence of categorical

perception.

Spratling & Johnson (2006) propose a computational model of perceptual learning processes and categorical perception. The authors build a bio-grounded architecture showing a functional differentiation between computations in apical dentrites (top-down feedback-dependent inputs from other regions, e.g. linguistic or attention processes) and basal dendrites (bottom-up sensory-driven inputs from sensors). Emulating the inter-cortical interaction, the unsupervised learning occurs at different stages of visual hierarchy and leads to the emergence of a categorical perception effect. The model shares with my proposal the idea that categorical perception is supported by an integration between bottom-up signals (input-driven) and top-down signals (feedback-driven). However, this model supports this integration trough a bio-plausible hardwired connectivity while my proposal exploits a novel learning rule emulating the integration of associative and reward-based signals in the brain (Caligiore et al., 2019b)

Kröger et al. (2007) propose a model of speech production showing a categorical effect. The model shows a neuro-inspired system-level architecture that includes many cortical and subcortical modules (e.g., sensory, motor and linguistic layers) and it is trained trough an unsupervised learning rule (self-organising maps; SOMs). Similar to mine, this model adopts a system-level modelling approach that aims to emulate many cortical and sub-cortical functions. However, the proposal adopts a pure unsupervised learning rule to train the weights between the layers while my proposal involves both unsupervised and reinforcement learning mechanisms. This allows me to better investigate how task demands affect the organisation of internal representations.

Salminen et al. (2009) propose a computational model that emulates the acquisition of categorical perception in infant human auditory systems. In particular, they produce many ecological inputs (vowel sounds) and adopt a bio-plausible hebbian SOM (unsupervised learning; UL). As in my work, the authors used realistic inputs to emulate the sensory processes. This solution improves the interpretability of the internal representations on the basis of more 'ecological

features' (e.g. vowel sounds or RGB pixels). However, the authors manipulate the input pattern frequencies to bias the SOMs for inducing the representation of prototypical categories. Instead, I used a set of input patterns with the same frequency and the model nevertheless acquires the representation of prototypical categories. Moreover, the authors adopt a pure UL rule while my model is trained with a novel rule that integrate both UL and RL.

Casey & Sowden (2012) propose a bio-plausible model reproducing the emergence of categorical perception in the brain visual system. The system is composed of three sequential layers, of which the first encodes low level-level visual features and the last receives both from the previous ones and from an external top-down source. This last top-down input causes the category learning. Each layer implements a competitive mechanism based on lateral inhibition and the whole architecture learns trough a bio-plausible unsupervised Hebbian learning rule. Similarly to my work, this model proposes a hierarchical visual system (composed of different sequential computation levels) and adopts ecological inputs to train the model. However, the model emulates the visual hierarchy abstracting other brain structures, namely the top-down feedback input is completely abstract while in my model it depends on many modules of a motivational system. Moreover, the model exploits an unsupervised learning rule while mine considers also reinforcement learning to encode the feedback. Last, the feedback mechanisms of the model only influence the top-layer while in my case the RL-based feedback biases both the top motor layer and the intermediate perceptual level.

Tajima et al. (2016) propose a computational model that emulates the neural populations dynamics during the acquisition of colour-based categorical perception. The model is supported by a simple recurrent neural network composed of a sensory and a category layer, in which a Bayesian inferential top-down process allows the second layer to influence the lower one in a categorical way. Despite this proposal adopts a neuro-inspired approach, it shows marked differences with respect to my work. The model does not use a true 'learning process', in that the emergence of categorical perception is based on a top-down inferential process.

In this sense, the model has common features with the first two computational studies, in which a recurrent neural network biases a sensory system and leads to the emergence of categorical perception. Moreover, the inferential process is not influenced by a performance-related feedback signal. Indeed, the model does not emulates the contribution of reward signal produced by subcortical structures as my model did.

Pérez-Gay et al. (2017) propose a computational model of categorical perception in which they investigate the underpinning different learning processes. They adopt a functional approach based on machine learning techniques, including an auto-encoder (AE; Goodfellow et al., 2017) and a classifier. The model undergoes an UL phase (only VAE training) and a supervised learning phase in which the whole model (both the trained VAE and the classifier) has to categorise the input on the basis of external labels. By comparing the internal representations of the AE after the UL phase with those after the SL phase, the authors detect a categorical perception effect. Similarly to my model, the authors adopt a functional approach based on a generative model and a classifier model. Moreover, they investigate how the interaction between an unsupervised and feedback-dependent phases can support the emergence of categorical perception. However, I adopt a neuro-inspired approach to build the model, showing a higher biological plausibility. Furthermore, my learning protocol involves a pure unsupervised learning phase only before the task start, while the task performance integrates both unsupervised and feedback-dependent signals.

Table 3.13 shows a list of the models I have taken in consideration here. The table highlights that most models encompass learning processes, with the exclusion of Beer (2003) and Tajima et al. (2016) involving evolutionary and inferential processes respectively. Moreover, several models adopt unsupervised learning rules. Despite unsupervised associative mechanisms have a key role in categorical perception, empirical evidence strongly points to the fact that several brain areas integrate multiple learning processes (i.e., supervised, unsupervised, reinforcement; Caligiore et al., 2019b). Interestingly no model on categorical perception

integrates reinforcement learning mechanisms, while my proposal shows both an unsupervised phase and an integrated unsupervised/reinforcement phase. At last, with the exclusion of Kröger et al. (2007) and my proposal, the bio-plausible models tend to focus on a particular brain system while abstracting the computations of other structures (system-level approach).

Limits and future directions

Although the previous section shows the advancements of my model with respect to the others, it still has limitations I intend to overcome in my future works, together with the development of interesting aspects. I discuss the main ones in this section.

Bio-plausibility and neuro-inspired/bio-grounded approaches. my computational proposal is supported by a neuro-inspired architecture in which the key components are implemented with neural networks (e.g., a generative neural network and an actor-critic network). Despite the architecture is functionally inspired by the interaction between cortical and subcortical brain systems (sensory-motor cortices and basal ganglia) and maintains a certain level of bio-plausibility (e.g., localistic learning rules; Illing et al., 2019), I used simplified neurons and abstract plasticity rules. Future work could aim to develop the ideas proposed here (overall architecture and systems interaction) based on neural networks having a higher degree of bio-logical detail. For example, I could build models based on spiking neurons and bio-grounded learning rules such as STDP (Zenke et al., 2015; Zappacosta et al., 2018), integrating plasticity rules that involve a reward signal as done in Rougier et al. (2005). Moreover, I could use spiking generative models (Neftci et al., 2014; Dasgupta & Osogami, 2016; Basanisi et al., 2020) to emulate the STDP effects on representation learning processes. These implementations would support further investigations about brain plasticity and the emergence of categorical perception.

Data fitting and model updates. my architecture shows perceptual processes that I qualitative compare with experimental evidence in healthy and clinical conditions. However, the computational tests presented here are only a 'proof-of-concept' as I need to test my model against detailed experimental data. To overcome this limitation, I aim to enhance the motor component of the model, for now representing a simplified output, making it able to produce performances comparable to those of humans.

Embodiment and robotic environment. I aim to follow a second complementary direction of neuro/cognitive robotics by linking the whole architecture to a robotic arm. This approach would allow the reproduction of human motor movements during a sorting task, supporting the investigation of cognitive processes underlying category learning. A robotic arm would allow the architecture to autonomously develop more complex embodied processes. For now the model emulates only some essential elements of embodiment (e.g. realistic sensory input; an environment feedback, based on the model performance, that influences the model perception) but a simulated or physical robotic environment would support deeper investigations on the relationships between categorical perception, motor skills and embodiment (Collins & Olson, 2014; Schendan & Ganis, 2012; Davis & MacNeilage, 2000), also in case of clinical conditions (e.g., autism; Taffoni et al., 2019).

Transfer learning skills and generalisation analysis. Here I consider three category learning tasks in isolation and a different model solves each one of the three task conditions (either sorting rule for colour, or shape, or size). I adopted this strategy due to the large amount of computational resources required to systematically study the multiple learning conditions (i.e., three sorting rules, five RL levels, two resource levels of perceptual component). In particular, I repeated the task for each of the thirty conditions for a total of over 1000 simulations. This approach allowed the execution of robust statistical analyses but it prevents testing the usefulness of representations acquired in a single task condition for the solution of

the other two task conditions (i.e., transfer learning; Canini et al., 2010; Shao et al., 2014). Moreover, my approach prevents the investigation of adaptive categorical perception, for which the model is required to further adapt its perception in case the sorting rule unexpectedly changes during the task performance. To overcome these issues, I aim to test the model with two further task conditions. First, I could implement a 'static generalisation condition' in which the model is tested with other categories after the 'principal task', keeping fixed the perceptual component. This test should clarify the relationship between the UL/RL balance and the generalisation skills of the perceptual component. Second, I could implement an 'adaptive categorical perception condition' in which the sorting rule suddenly changes many times and the model has to online adapt its perception and response to the new requests. This test should clarify the relationship between the UL/RL balance and the perceptual adaptation of the model. Overall, I expect that the extreme RL model (L4), producing the most task-directed representations, might lose the generalisation and adaptation capacities due to its extreme information loss impacting on the task-independent features. Conversely, the models with a more balanced RL/UL ratio (L2 and L3), could show the best performances both in the main sorting task (as shown here) and in these two new tasks. This would corroborate the idea that a balanced UL/RL mix is the most suitable solution for an artificial and biological perceptual component, needing to adapt to an uncertain environment where the task can change (e.g. novel objects to categorise) and the computational resources are limited.

Multi rules categorisation and catastrophic forgetting. The model is able to adapt its motor, motivational, and perceptual components to solve a sorting task that shows a fixed single sorting rule (sort for colour, shape, or size). Although the system could slowly adapt itself after a rule change, it would likely incur into catastrophic forgetting (i.e. the loss of the already acquired information caused by the acquisition of new ones; McCloskey & Cohen, 1989; Knoblauch et al., 2014). This limit is strongly linked to the previous ones, due to the fact that an ideal perceptual system should be able (1) to transfer the knowledge to another task

or task condition (transfer learning) without losing the previously acquired information (catastrophic forgetting) and (2) to quickly adapt itself in case the initial sorting rule changes (adaptive categorical perception). To overcome this limitation I could integrate the architecture presented here with mechanisms implementing an internal manipulation of perceptual representations as studied in the first two computational models. In those models, a dynamical working memory encodes different categorisation rules and guides an internal 'top-down manipulator' that selects different portions of a visual neural network. The integration of this internal manipulation and the learning processes studied here should allow an architecture to select and train specific portions of a neural network, improving the problem of catastrophic interference and quick perceptual adaptation (e.g., 'experts approach'; Tommasino et al., 2019).

Category learning, categorical perception and perceptual learning: differences and model updates. my proposal focuses on CP in case the task-directed actions (e.g. category learning) alter the perceptual representations (differences and similarities expansion). On the other hand, 'perceptual learning' refers to the 'experience-dependent enhancement of my ability to make sense of what I see, hear, feel, taste or smell' (Gold & Watanabe, 2010). Interestingly, Carvalho & Goldstone (2016) suggest that category learning and perceptual learning could share specific learning mechanisms, as in case of the emergence of categorical perception. Despite these commonalities, controversial evidences highlight some differences between these processes. For example, it is not clear if category learning and perceptual learning influence the perceptual systems at the same level (early, middle or late processing stages). To clarify the controversial evidences, I could extend my investigations executing specific model updates. For example, I could apply the same learning rule, integrating UL and RL, in each sensory-motor hierarchy of my networks. In particular, in addition to the second RBM of the DBN (from the first hidden layer to the second hidden layer of DBN), I could apply the same RL/UL rule on the first RBM (from the input layer to the first hidden layer of DBN). In this way I could potentially set different levels of UL/RL

integration at each level of abstraction (from the low-level perceptual processes to the motor selection). Searching for the model configurations that best fit the human data, I could investigate the differences in learning processes supporting category learning and perceptual learning, in particular the reward/task influence at different levels abstraction.

3.3.6 Conclusions

In this study I corroborated the motivated categorical perception hypothesis, for which the interaction between unsupervised and reinforcement learning leads to the emergence of human categorical perception. Integrating neuroscientific evidence and machine learning methods (e.g. generative neural networks), I built a neuro-inspired computational model that is able to perform a category learning task. In particular, the system-level architecture shows neuro-inspired components (emulating cortical and sub-cortical brain functional macro-systems) and integrates bio-plausible unsupervised and reinforcement learning processes (e.g., distributed representations and localistic learning rules). The analyses of internal representations and performance suggest that a balanced mix of unsupervised and reinforcement learning supports the acquisition of suitable task-directed representations (categorical perception), leading to the best performances. Instead, extreme cases lead to sub-optional performances due to maladaptive representation learning processes. In particular, in the case of limited computational resources the models without reinforcement learning are not able to focus on relevant features thus producing sub-optimal performances. Instead, the models without unsupervised learning show more unstable and slow learning processes, especially at early phases of learning, thus incurring in clustering errors and an excessive loss of information.

The model qualitatively reproduces experimental evidence in healthy condition, namely the emergence of category-based representations in extrastriate cortices. Moreover, the model can explain the altered categorisation performance in clinical conditions as autism. For example, the model with only unsupervised learn-

ing shows an excessive sensory autonomy with respect to the task-dependent feedback, possibly explaining the worse categorisation processes in some autism conditions. Moreover, the model with only reinforcement learning explains the low generalisation skills in some other autism conditions, due to an excessive loss of information. These opposite effects can explain the heterogeneity of autism spectrum conditions, as different imbalanced mix of unsupervised and supervised learning mechanisms in different autistic people.

The model could also support the development of machine learning systems able to undergo categorical perception effects, and robotic systems needing to face uncertain environments trough suitable representations. In particular, my neuro-inspired algorithm could prompt the development of new algorithms that are able to autonomously balance UL and RL processes depending on the task demands, the available computational resources, and generalisation requirements.

Chapter 4

Applications and theoretical advancements

4.1 The three-fold hypothesis and computational psychiatry: the case of autism spectrum condition

Here I report an application case in which I adopt the model 2 to investigate the role of inner speech in clinical populations of different ages. In addition to further corroborating the three-component hypothesis, this application case provides insights into the role of inner speech in the Autism Spectrum Condition (ASC).

In particular here I focus on mild autism, namely a neurodiversity condition that leads to repetitive behaviours, social impairment, sensory alterations, and restricted interests (Association et al., 2013). 'Mild autism' is a diagnostic label, previously corresponding to the 'Asperger syndrome'/'high-functioning autism', that involves low intensity symptoms compared to the other two severity levels ('moderate' and 'severe'). There is also an open debate regarding the use of the terms 'condition' or 'disorder' to refer to mild autism/Asperger Syndrome/high-functioning autism (Jaarsma & Welin, 2012). Here I use the term 'condition' to avoid stigma without ignoring the daily challenges and possible impairments it involves.

4.1.1 Theoretical premises and methodological approach

Several experimental and clinical studies started to investigate the role of inner speech in psychiatric and neurodiversity conditions (Petrolini et al., 2020). These studies show that inner speech can generally provide cognitive support, but in some cases it can also have disruptive effects. For example, in schizophrenic patients it can be distracting, fragmented, charged with negative emotions, and possibly involve auditory hallucinations.

Importantly for this computational study, many works have investigated the relationship between inner speech and executive functions in ASC, and contrasting results are reported (for a review see Williams et al., 2016). For example, some studies on planning (Wallace et al., 2009; Williams et al., 2012; Holland & Low, 2010) found that an experimental interference of inner speech (e.g., articulatory suppression) impairs planning abilities in control participants but not in an ASC population. However, the results should be taken with caution due to potential methodological limitations (see critiques by Williams et al., 2016). Again, evidence on working memory suggests that ASC individuals do not spontaneously use inner speech to name stimuli internally (e.g., Joseph et al., 2005) while most studies on motor control indicate either that ASC individuals use inner speech or that the absence of inner speech does not impact their performance. Crucially for us, Russell-Smith et al. (2014) showed that articulatory suppression does not interfere with cognitive flexibility in ASC people, who do not show an impaired performance. However, Winsler et al. (2007) found that ASC children performed worse than controls even if they did used private speech. Overall, these scattered and controversial results leave space to further research. In particular, findings suggest that autistic people make a reduced use of inner speech but it is debated whether this reduction has an impact on executive functioning and in particular on cognitive flexibility.

Here I have used the model 2 (see section 3.1 and section 3.2 for a complete description of this model) to investigate the relationship between the inner speech

and executive functions in Autism. In particular, I have taken in consideration four already published and validated experimental studies in which it is administrated the WCST to eight control and autistic human populations. We focus on these studies because they (a) adopted the Heaton's version of WCST, (b) involved an ASC group (mild autism/Asperger syndrome/high-functioning autism) and a matched control group, and (c) reported at least CC, PE, and NPE indices. The first group (Shu et al., 2001) involved 26 children (6 to 12 years) with a diagnosis of autism without mental retardation (DSM-III) and a control group of 52 children matched for age. The second group (Kaland et al., 2008) involved 13 teenagers (16.40 ± 2.84) with a diagnosis of Asperger syndrome or High-functioning Autism (ICD-10) and a control group of 13 teenagers matched for age and QI. The third group (Rumsey, 1985) involves 9 young adults (27 ± 7) with a diagnosis of autism without mental retardation and high verbal competences (DSM-III) and a control group of 10 young adults matched for age, education and QI. The fourth group (Ambery et al., 2006) involves 27 old adults (33.5 ± 12) with a diagnosis of Asperger syndrome (ICD-10) and a control group of 20 old adults matched for age and QI. Despite the populations of Rumsey (1985) and Ambery et al. (2006) show similar ages (27 ± 7 vs 33.5 ± 12) I define them 'young adults' and 'old adults' to better distinguish them. I did not found studies that administrate the WCST to ASC adults older than such age.

4.1.2 Results

Configurations of parameters of the best fitting models

As done in the first and second computational studies (e.g., section 3.1.4 and section 3.2.3), I have used a statistical search method based on the minimisation of the mean square error (MSE) to find the models parameters. In particular, this method is suitable to find the parameter configurations that best reproduced the behavioural data of the control and ASC populations. Although the sample size of some groups is small, the model reproduces the human behavioural data with a

low average MSE for both control and ASC groups.

Table 4.1 and Figure 4.1 show the parameter values of the models populations that best fit the dataset of the human groups. The parameters represent the the simulated cognitive traits of the model and, therefore, of the modelled human participants. Regarding the inner speech contribution (parameter λ), the control groups show an increasing tendency depending on ageing. Differently, ASC groups show an absent or negligible inner speech contribution in all ages.

Regarding the error sensitivity (parameter μ), the control groups show an "inverse U-shaped" curve. In particular, children and old adults show a similar and lower error sensitivity, while teenagers and young adults show a similar and higher error sensitivity. In the case of ASC groups, I found similarities among pairs of different groups. In particular, children and young adults show a similar and lower error sensitivity, while teenagers and old adults show the same higher error sensitivity.

Regarding the memory refresh/forgetting speed (parameter ϕ), the control groups again show similarities between children and old adults. Differently, teenagers show the lowest value and young adults the highest value. In case of ASC groups, I found a descending tendency. In particular, children show the highest value with respect to the other groups, and the latter ones show similar values.

Regarding the distractibility/exploratory behaviour (parameter τ), the control groups have similar values. Despite this, children and old adults show the same slightly higher value with respect to teenagers and young adults, that show the same value. In the case of ASC groups, similarly to the ϕ parameter, I found a descendent tendency. In particular, children show higher value with respect to the other groups, and the latter ones are similar between them.

Behavioural comparisons

Comparisons between perseverative errors and non perseverative errors in each group Since perseverative errors and non perseverative errors identify two opposite tendencies, respectively for perseveration and for distraction (for more

	Error sensitivity (μ)	Memory refresh, Forgetting speed (ϕ)	Distractibility, Explorative behaviour (τ)	Inner speech contribution (λ)
Control models				
Children	0.08	0.37	0.18	0.17
Teenagers	0.17	0.09	0.12	0.23
Young adults	0.21	0.73	0.12	0.33
Old adults	0.05	0.41	0.18	0.52
ASC models				
Children	0.11	0.93	0.83	0.01
Teenagers	0.20	0.19	0.14	0.0
Young adults	0.08	0.11	0.08	0.02
Old adults	0.20	0.19	0.14	0.0

Table 4.1: Values of the parameters of the models that produce the best fit of the data on the WCST indices.

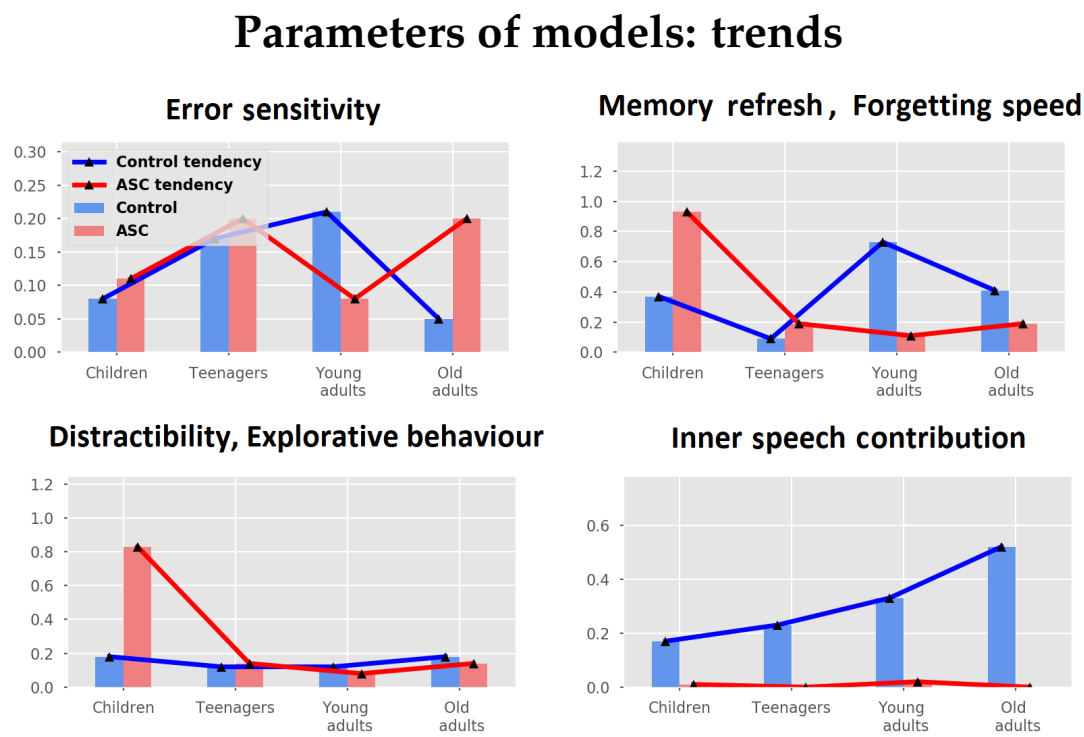


Figure 4.1: Graphic visualisation of the parameters of the models that best fit the datasets of the human groups (Children, Teenagers, Young adults, Old adults).

details on this interpretations see section 3.1.5, I performed statistical comparisons (t-tests with Bonferroni's correction) between PEs and NPEs of each model to investigate its behavioural profile (Figure 4.2).

The results show that in the control condition only old adults have significant

Behavioural indices of models (intra-condition analysis: PE vs. NPE)

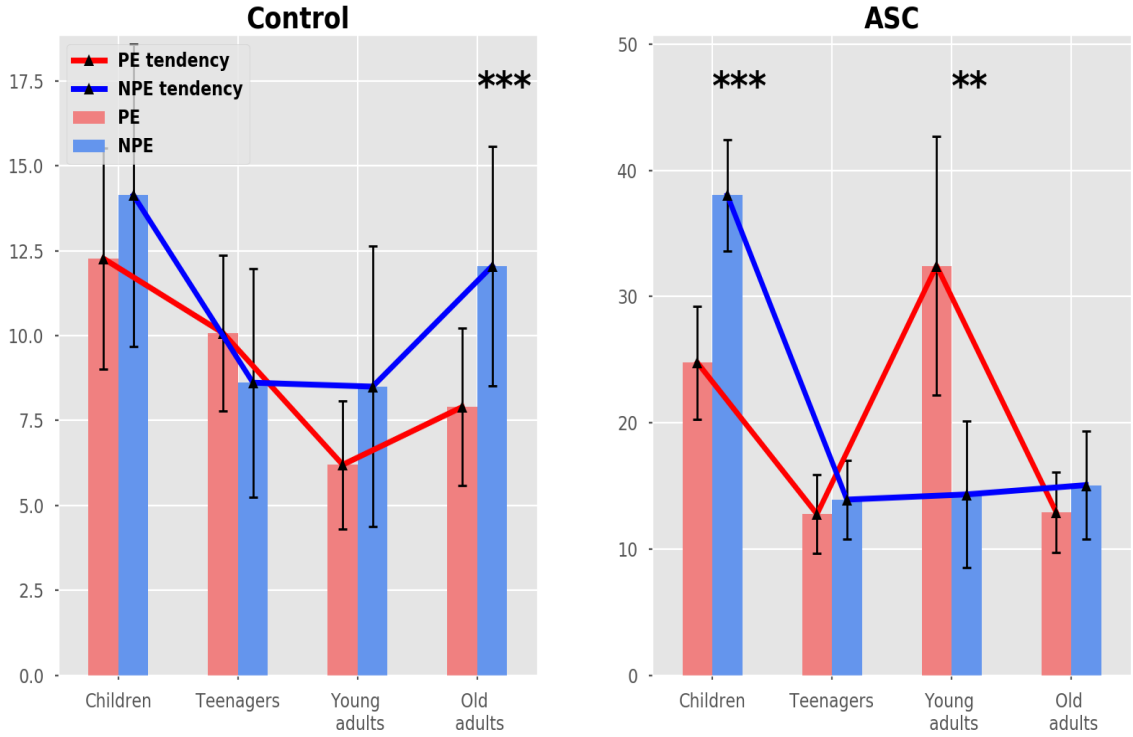


Figure 4.2: Comparisons between PE and NPE in the control and ASC conditions (Children, Teenagers, Young adults, Old adults).

differences in their behavioural profile, with an imbalance toward NPE (7.9 ± 2.32 vs 12.05 ± 3.53, $p < .001$). In the ASC condition, I found that children have an imbalance toward NPE (24.77 ± 4.48 vs 38.04 ± 4.4, $p < .001$) while young adults have an imbalance toward PE (32.44 ± 10.23 vs 14.33 ± 5.79, $p < .01$). Despite the plots show many imbalances of PE and NPE population means in the other groups of models, they also show a high population variability that prevents further statistical differences.

Comparison between the behaviour of different age groups (intra-condition analysis) I performed statistical comparisons (one-way Anova and post-hoc t-tests with Bonferroni's correction) between the models of each condition. These analyses aimed to investigate the differences in the ageing process of control and ASC conditions (Figure 4.3, blue and red trend lines).

Behavioural indices of models (inter-conditions analysis: control vs. ASC)

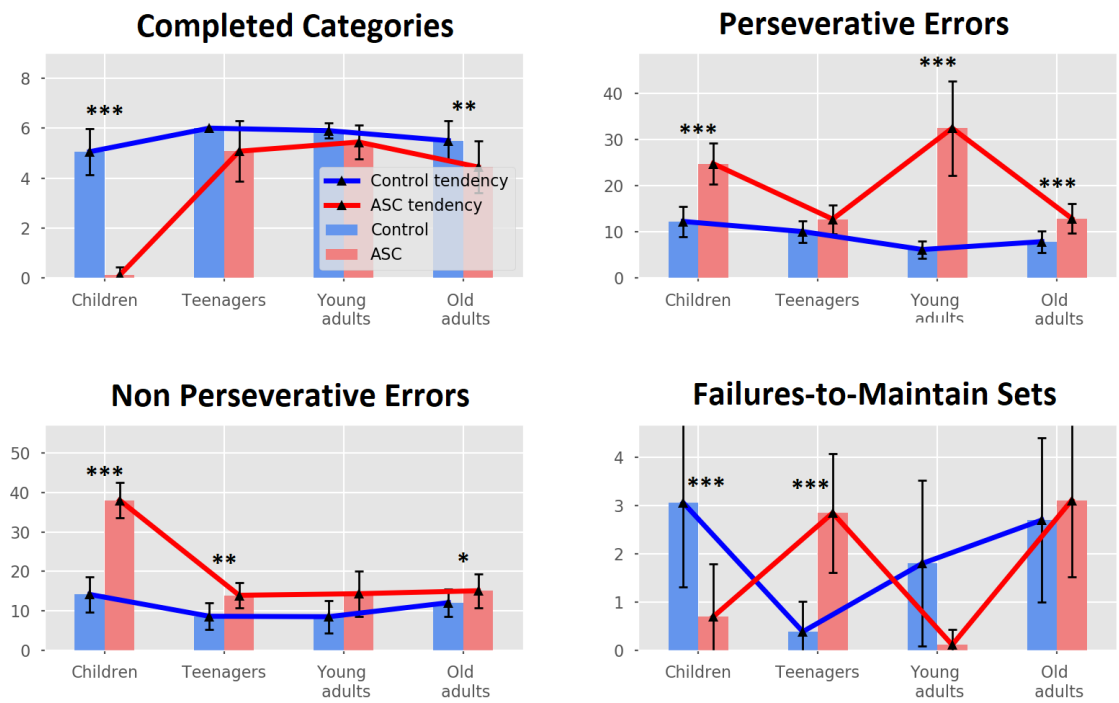


Figure 4.3: Behavioural indices and comparisons of all models (Children, Teenagers, Young adults, Old adults).

Regarding the completed categories index (CC), I found statistical difference between the control models ($F = 7.03$, $p < .001$). Post hoc tests indicate that children achieve a lower CC index with respect to teenagers (5.06 ± 0.93 vs 6.0 ± 0.0, $p < .001$). I did not find significant statistical differences between the other models, probably due to the high variability of each model population. I found statistical difference between the ASC models ($F > 50$, $p < .001$). Post hoc tests indicate that children achieve a low CC index with respect to teenagers (0.12 ± 0.32 vs 5.08 ± 1.21, $p < .001$), young adults (0.12 ± 0.32 vs 5.44 ± 0.68, $p < .001$), and old adults (0.12 ± 0.32 vs 4.44 ± 1.03, $p < .001$). I did not find further significant statistical differences between the other models.

Regarding perseverative errors (PE), I found statistical difference between the control models ($F = 19.87$, $p < .001$). Post hoc tests indicate that children have high PE with respect to young adults (12.27 ± 3.26 vs 6.2 ± 1.89, $p < .001$) and old adults

(12.27 ± 3.26 vs 7.9 ± 2.32, $p < .001$). I did not find further significant statistical differences between the other models. I found statistical difference between ASC models ($F > 50$, $p < .001$). Post hoc tests of ASC models indicate that children show higher PE with respect to teenagers (24.77 ± 4.48 vs 12.77 ± 3.12, $p < .001$) and old adults (24.77 ± 4.48 vs 12.93 ± 3.17, $p < .001$), and lower PE with respect to young adults (24.77 ± 4.48 vs 32.44 ± 10.23, $p < .001$). I did not find further significant statistical differences between teenagers and young adults.

Regarding non perseverative errors (NPE), I found statistical difference between control models ($F = 9.82$, $p < .001$). Post hoc tests indicate that children have high NPE with respect to teenagers (14.13 ± 4.44 vs 8.62 ± 3.36, $p < .001$) and young adults (14.13 ± 4.44 vs 8.5 ± 4.13, $p < .01$). I did not find further significant statistical differences between the other models. I found statistical difference between ASC models ($F > 50$, $p < .001$). Post hoc tests of ASC models indicate that children have higher NPE with respect to teenagers (38.04 ± 4.4 vs 13.92 ± 3.12, $p < .001$), young adults (38.04 ± 4.4 vs 14.33 ± 5.79, $p < .001$), and old adults (38.04 ± 4.4 vs 15.07 ± 4.29, $p < .001$). I did not find significant statistical differences between the other models.

Regarding failure-to-maintain sets errors (FMS), I found statistical difference between control models ($F = 10.04$, $p < .001$). Post hoc tests indicate that children have high FMS with respect to teenagers (3.06 ± 1.75 vs 0.38 ± 0.62, $p < .001$), and old adults have higher FMS with respect to teenagers (2.7 ± 1.71 vs 0.38 ± 0.62, $p < .01$). I did not find significant statistical differences between the other models. I found statistical difference between ASC models ($F = 24.31$, $p < .001$). Post hoc tests of ASC models indicate that children have lower FMS with respect to teenagers (0.69 ± 1.1 vs 2.85 ± 1.23, $p < .001$) and old adults (0.69 ± 1.1 vs 3.11 ± 1.59, $p < .001$). Moreover, teenagers have higher FMS with respect to young adults (2.85 ± 1.23 vs 0.11 ± 0.31, $p < .001$), and young adults have lower FMS with respect to old adults (0.11 ± 0.31 vs 3.11 ± 1.59, $p < .001$). I did not find significant statistical differences between the other models.

Comparison between the behaviour of the control and experimental groups (inter-condition analysis) I performed statistical comparisons (t-tests with Bonferroni's correction) between the indices of the control and ASC models to investigate the behavioural differences between them in each age (figure 4.3)

Regarding the completed categories (CC), I found that they are lower in ASC children (5.06 ± 0.93 vs 0.12 ± 0.32, $p < .001$) and ASC older adults (5.5 ± 0.81 vs 4.44 ± 1.03, $p < .01$). I did not find a statistical differences in teenagers (6.0 ± 0.0 vs 5.08 ± 1.21, $p > .05$) and young adults (5.9 ± 0.3 vs 5.44 ± 0.68, $p > .05$).

Regarding perseverative errors (PE), I found that they are higher in ASC children (12.27 ± 3.26 vs 24.77 ± 4.48, $p < .001$), ASC young adults (6.2 ± 1.89 vs 32.44 ± 10.23, $p < .001$), and ASC old adults (7.9 ± 2.32 vs 12.93 ± 3.17, $p < .001$). I did not find any statistical difference in teenagers (10.08 ± 2.3 vs 12.77 ± 3.12, $p > .05$).

Regarding non perseverative errors (NPE), I found that they are higher in ASC children (14.13 ± 4.44 vs 38.04 ± 4.4, $p < .001$) and in ASC teenagers (8.62 ± 3.36 vs 13.92 ± 3.12, $p < .01$). I did not find any statistical difference in young adults (8.5 ± 4.13 vs 14.33 ± 5.79, $p > .05$) but I found a slightly higher value in ASC old adults (12.05 ± 3.53 vs 15.07 ± 4.29, $p < .05$).

Regarding the failure-to-maintain set errors (FMS), I found that these are lower in ASC children (3.06 ± 1.75 vs 0.69 ± 1.1, $p < .001$) and higher in ASC teenagers (0.38 ± 0.62 vs 2.85 ± 1.23, $p < .001$). I did not find any statistical difference in young adults (1.85 ± 1.72 vs 0.11 ± 0.31, $p > .05$) and old adults (2.7 ± 1.71 vs 3.11 ± 1.59, $p > .05$).

Internal functioning comparisons

I also investigated the internal functioning of the models. Figure 4.4 shows the internal activation of the working memory units of the models recorded during their task performance. The activation of each unit corresponds to a specific sorting rule to follow and the top-space of each plot of the figure shows the errors that occur during each card response.

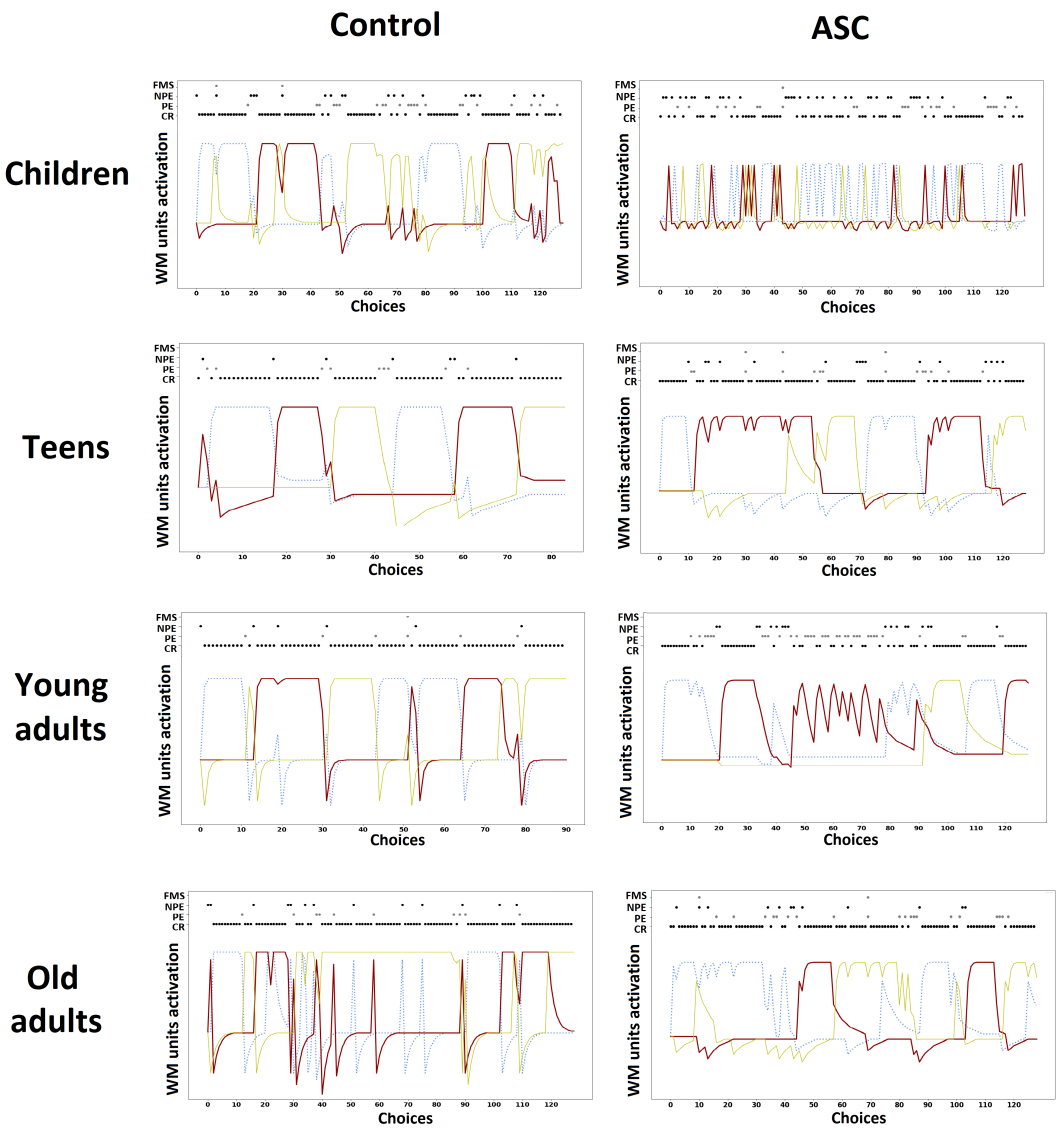


Figure 4.4: Internal functioning of the executive working memory of the control and ASC models. Each line represents the activation of a memory unit encoding a specific matching rule: thick red line: colour-based matching rule; dotted thin blue line: shape-based matching rule; continuous yellow line: size-based matching rule. The dots at the top of graphs indicate the instances of correct responses (CR) or errors (PE, NPE, FMS).

In the case of children, the activation of the working-memory units of the control and ASC models appear very different. In particular, the ASC model shows several erratic strategy changes that cause the occurrence of several NPE. Interestingly, despite the model is evidently distracted and does not keep the focus on a specific strategy, few PE are scored. As already shown in section 3.1.5 and section 3.2.4, a participant with high distractability can choose by chance an already tried strategy thus erroneously appearing perseverative. Here I refer to these errors

as 'distraction-related PEs'. At last, also the control model shows a sub-optimal performance, caused by reasoning errors (e.g., see the 65-80 interval of trials) and attention failures (e.g. 3-4, 25-35 interval)

In the case of teenagers, the control model shows a good landscape with some negligible reasoning failures (e.g., 0-5 interval) and perseveration (e.g., 40-45 interval). The ASC model shows several 'sustained attentional failures' (e.g., 10-40 interval) and reasoning errors (e.g. 110-120 interval) that cause many NPE and FMS errors.

In the case of young adults, the control model shows a good landscape with minor attention failures (e.g. 15-20, 50-55 intervals). The ASC model shows many perseverative behaviours (e.g. 40-70 interval) and attention failures (e.g. 75-85 interval).

In the case of old adults, both control and ASC models show sub-optimal landscapes. The ASC model shows many attention failures (e.g., 35-45 interval) and reasoning errors (e.g., 55-65 interval), causing many PEs and NPEs. Interestingly, the control model shows many FMS (e.g., 50-75 interval) as the ASC model, showing a poorly focused behaviour.

4.1.3 Discussion and conclusions

The model 2 reproduces most behavioural indices of control and autistic groups performing the Wisconsin Cards Sorting Test. Moreover, it captures several intra-group and inter-group cognitive and behavioural differences.

Regarding control populations, I generally found similar parameters values between children and old adults (Figure 4.1, blue lines) in error sensitivity, memory refresh/forgetting speed, and distractibility/exploratory behaviour, detecting some 'U-shaped tendencies' related to age. Differently from the other three parameters, I found an inner-speech contribution that increases with age, being low in children and high in adults. Further investigations of the cognitive profile of control groups confirmed the U-shaped trends in perseverative errors (perseverative

behaviour) and non-perseverative errors (attention/reasoning failures) (Figure 4.2, left plot). At last, a qualitative analysis of internal activations of the models corroborated these trends (Figure 4.4), showing that teenagers and young adults exhibit the best performance with respect to children and old adults in which I found more sub-optimal behaviours affected by distraction and perseveration.

Despite the emergence of these trends, the cognitive differences (parameters) between controls groups do not always cause statistically significant differences in behavioural data (Figure 4.3). For example, only children show significantly lower global performances than the other groups (teenagers, young adults, and old adults), which do not show statistical differences between them.

These results allow the interpretation of contrasting findings on ageing-related effects. In particular, several studies indicate that ageing causes significant brain changes (e.g., Sullivan et al., 2001; Peters, 2006), in particular a weakening of executive functions (Cepeda et al., 2001; Fisk & Sharp, 2004; Samanez-Larkin & Knutson, 2015), but other studies reveal compensating brain processes such as functional reorganisation and increased bilateral recruitment (Cabeza & Dennis, 2012; Daselaar et al., 2013).

Considering this literature, based on the results presented here I suggest that the inner speech contribution, showing an increasing trend from children to old adults, can play an ageing compensation effect. In particular, I propose that inner speech contributes to *support early development* and to *avoid/compensate cognitive decline*, thus mitigating the life-span cognitive and behavioural differences between neurotypical individuals. This proposal is also coherent with my results from the second computational study (section 3.2.4), highlighting that inner speech interacts with the other cognitive processes (working memory storing, error sensitivity, attention), boosting the global performance and diminishing distracted and perseverative behaviours. Moreover, my proposal corroborates the several studies that highlight an important executive modulator function of inner speech in old adults (e.g., Kray et al., 2004; Fry, 1992; John-Steiner, 2014).

Regarding the ASC populations, I found relevant differences in cognitive profiles with respect to the control populations. First, I found that ASC groups do show a reduced contribution of inner speech along the life-span. Second, I found greater differences between children and other groups regarding working memory decay and distractibility. Third, autistic groups show different imbalances with respect to control groups (figure 4.2, right plot). In particular, autistic children show an evident imbalance toward distractibility (NPE), while young adults show an imbalance toward perseverative behaviours (PE).

These results are particularly interesting because the diagnostic criteria for autism rely on repetitive behaviours (Association et al., 2013) and clinical studies mostly focus on perseverative/repetitive behaviours in ASC children and adults (Carcani-Rathwell et al., 2006; Lopez et al., 2005). On the other side, several works have investigated attention abnormalities in autism suggesting that an attention impairment could play a causal role in the development of ASC individuals (for a review see Keehn et al., 2013). The results presented here agree with these last studies, suggesting that ASC children mostly show an imbalance toward distractions with respect to perseverative behaviours. Moreover, the models suggest a cognitive change in ASC peoples along the life-span, from a distracted profile in children to a perseverative one in young adults.

Regarding behavioural age-related differences, the cognitive traits (parameters) seem to have a more marked effect on behaviours in ASC peoples with respect to the control groups. For example, the descending values of distractibility and memory refresh are reflected by the similar curve of NPE and the low error sensitivity in children and young adults cause higher PE with respect to teenagers and old adults. However, in the case of children this result is evidently altered by many distractibility-related PEs. In particular, the extreme distractibility of ASC children causes a random behaviour (Figure 4.4, first row) that is sometime scored as 'perseverative behaviour' although it is caused by attention failures (see the imbalance toward NPEs in Figure 4.2, right plot).

Interestingly, the FMS curve shows a different and unexpected trend with respect

to the control trends. In particular, I could expect that ASC children would show higher FMS due high distraction, but in fact they showed a low value of this index. This is probably explained by the difference between NPE and FMS, where the first indicate an attentional/reasoning failure and the second indicates a sustained attention failure. Since ASC children cannot focus on a specific strategy (sorting rule) for long, they often do not achieve the necessary number of responses to occur in a sustained attention error (FMS). These results are coherent with Sinzig et al. (2008), detecting an impairment in selective attention and not sustained attention, and with Johnson et al. (2007), detecting a response inhibition impairment rather than a sustained attention impairment. A high FMS in old adults is another interesting data. While a higher FMS value of teenagers is expected and corroborates a sustained attention deficit (Christakou et al., 2013; Murphy et al., 2014), I could expect an imbalance toward perseveration in old adults. Instead, I found higher FMS with respect to young adults without a marked PE/NPE imbalance (Figure 4.2, right plot). These results need further investigations, in particular regarding sustained attention in autistic old adults.

In summary, comparing control and ASC populations I found statistically lower performance only in ASC children and ASC old adults with respect to their control groups. The behaviour comparisons (Figure 4.3) and the analysis of the internal activation of the models (Figure 4.4) suggests that these differences are caused mainly by more distractions in ASC children/teenagers and a higher perseveration in ASC young/old adults. These results suggest an immature executive functioning in ASC children and a slight cognitive decline in old adults, as suggested by similar trends in the control groups. Despite this, the control groups show weaker intra-condition behavioural differences than the ASC groups, where the difference between age groups appears more marked.

Although many latent variables can contribute to these different behavioural performances (e.g., impaired social learning in autistic children, for a review see Tomasello et al., 1993), my data suggest that the lack of inner speech development in ASC people could make the ageing effects more evident. In particular, since in

control conditions the inner speech represents a cognitive support for an immature executive functioning in children and a compensating process in old adults, its absence in autistic peoples could deprive them of these compensation processes.

My hypothesis can contribute to explain the contrasting evidence of studies on autism, inner speech and executive functions (for a review see Williams et al., 2016). In particular, the differences might be due to the heterogeneous involved populations that span from children to old adults. Moreover, this proposal is coherent with many studies regarding autism and life-span cognitive changes (Happé et al., 2006; Pellicano, 2010) suggesting that also autistic people show an improvement of executive functions during the life-span. Indeed, in ASC teenagers and young adults, compensating processes emerge (e.g., higher visual skills and visual thinking with respect to neurotypical peoples; Bókkon et al., 2013; Grandin, 2009; Mottron et al., 2006). However, the lack of inner-speech support still represents a strong impairment for children and old adults.

Finally, my results have interesting clinical implications. Many therapeutic approaches aim to limit compromising symptoms in autism (Aman, 2005), but only few of them focus on speech abilities in autism (Adams et al., 2012; Fernandes et al., 2012; Flippin & Hahs-Vaughn, 2020). These approaches aim to increase linguistic skills to improve social communication abilities, but they do not directly focus on self-directed language (inner-speech). This study suggests that clinicians should device a new class of therapeutic approaches primarily focusing on developing inner speech skills in autistic children. In particular, the integration of early development of inner speech and strong visual thinking could represent an important cognitive support along the life-span of autistic people, from childhood to adulthood.

4.1.4 Limitations and future directions

Although this work successfully integrates participatory research with computational modelling for clinical scopes, it shows limitations that I aim to overcome in

our future works.

First, the experimental studies I have considered (Shu et al., 2001; Kaland et al., 2008; Rumsey, 1985; Ambery et al., 2006) do not include a verbal shadowing protocol that directly evaluates the inner-speech contribution. Despite this, the model has already demonstrated to disentangle the inner-speech contribution during an experimental protocol that integrates the WCST with a verbal shadowing protocol (section 3.2). Here I have exploited the model to propose inferences about individual differences, and the predictions are compatible with the proposals of the literature, also producing useful interpretations for clinicians. Future experimental studies with autistic people should integrate the administration of WCST with a verbal shadowing as done in Baldo et al. (2005). I will consider future experimental data with the aim of testing the model's predictions and building a comprehensive theory of experimental, neuropsychological, and computational aspects of Autism.

Second, the sample size of groups extracted from Kaland et al. (2008) and Rumsey (1985) is small. Although I have detected inter-groups statistical differences (e.g., more PEs in autistic young adults or more NPEs in autistic teenagers), this factor could lead to no difference in global performances in these groups. Future experimental studies should focus on the aim to enlarge the sample size of experimental groups, making more robust the interpretations regarding inner-speech and autism.

Third, the age difference between young adults (27 ± 7) and old adults (33.5 ± 12) is not large and it could alter my results. This point represents a general lack of literature on autistic people, in fact I have not found studies administrating the WCST to autistic adults older than such age. New experimental studies should aim to cover a wider age range, especially towards autistic old adults. This could get clinicians to consider Autism as a life-span condition, from childhood to old age.

4.2 The three-fold hypothesis and conscious processing: from flexible goal-directed behaviour to consciousness

In this section I propose a second 'application case' of my key theoretical proposals. In particular, I show how my key theories on flexible goal-directed behaviour can be extended toward the investigation of consciousness. Departing from the three-component hypothesis, I propose a new theory that describes the relationship between representations manipulation and conscious processing.

Note that this section is highly speculative, however it is built on the basis of my corroborated previous proposals and on the main theories of consciousness.

4.2.1 Background and theoretical premises

For centuries consciousness has been an hotly debated question in philosophy (e.g., Chalmers, 1995; Dennett, 2018). In recent decades, theoretical and technological advancements in cognitive neuroscience allow this topic to become a main target of scientific investigation. Many scientific theories propose a link between the brain and consciousness, focusing on several aspects such as the integration of information (Tononi, 2008; Tononi et al., 2016; Koch et al., 2016); the dynamic activation of central and peripheral cognitive/emotional systems of the brain (Damasio, 1989; Meyer & Damasio, 2009); the selection of relevant information at the central brain level and its 'broadcasting' to peripheral areas (Baars, 1997; Baars et al., 2003; Baars, 2005; Baars et al., 2013); the orchestration of the activation of multiple hierarchical brain systems by the frontoparietal system (Dehaene et al., 1998a; Dehaene & Naccache, 2001; Dehaene & Changeux, 2011); the reliance on higher-order representations that possibly involve the agent itself (Brown et al., 2019; Cleeremans, 2011); and the coordination of effective brain-body-environment sensorimotor interactions (O'Regan & Noe, 2001; O'Regan et al., 2005). On the

other end, advancements in computational modelling and cognitive robotics allow a better emulation of important cognitive processes relevant for consciousness. The advancement of interdisciplinary fields such as *Machine consciousness* (Reggia, 2013) and consciousness-inspired machine learning (e.g. Bengio, 2017) best exemplify the potential synergies between these scientific and technological approaches.

Notwithstanding the relevance of these advancements, both theories of consciousness and computational proposals still show limitations. First main theories of consciousness do not propose a systematic integration with studies on goal-directed behaviour (GDB). This lack prevents the emergence of a clear functional perspective on how consciousness guides important high level processes such as planning and problem solving. Second, many theories of consciousness lack an articulate neuro-computational perspective while modern theories of cognition consider the brain as a 'computational machine' (Churchland & Sejnowski, 1992; Dayan & Abbott, 2001). Although some theories have led to computational models of consciousness (e.g., Dehaene et al., 1998b; Pasquali et al., 2010; Tononi, 2008), there is still not a clear description of the system-level information manipulations that occur during a conscious state. Third, robotic systems still show a rigid behaviour failing to face novel goals and conditions and to exhibit a flexible general-purpose cognition (Hassabis et al., 2017; Lake et al., 2017). These limitations impact both the theories of consciousness and technological fields, requiring an efficient integration and synergies between them.

Overall, my research approach and theoretical proposals can be a starting points for facing these issues. In section 2.3 I introduced the three-component hypothesis, proposing that human flexible cognition and behaviour are supported by a goal-directed manipulation of internal representations. Departing from my previous proposals, I introduce here the *Representation Internal Manipulation* (RIM) neuro-computational theory of consciousness. This theoretical framework extends the three-components hypothesis, describing the computational processes at the basis of conscious flexible goal-directed cognition and behaviour. Indeed, the RIM theory can be considered (1) a four-components theory of flexible cognition, (2) a

formal model of goal-directed behaviour, and (3) a neuro-computational theory of consciousness.

4.2.2 The representations internal manipulation theory: a new four-components hypothesis of conscious flexible goal-directed behaviour

The theory rests on these key elements, now presented in detail: (1) an overall adaptive function ascribed to consciousness; (2) neural representations within the brain supporting conscious goal-directed processes; (3) four brain anatomo-functional macro-systems supporting such processes; (4) five classes of computational operations performed by conscious processes on internal representations.

The adaptive function of consciousness.

Some theories of consciousness propose an adaptive function of consciousness. In particular, the GWT (Baars, 1997; Baars et al., 2003; Baars, 2005; Baars et al., 2013) and GNWT (Dehaene et al., 1998a; Dehaene & Naccache, 2001; Dehaene & Changeux, 2011) suggest that it boosts decision making and flexible behaviour.

The RIM theory proposes that the brain architecture and processes supporting consciousness emerged due to the evolutionary opportunity to empower behavioural flexibility by enhancing the underlying goal-directed processes such as decision making, planning, and problem solving.

In particular, I posit that *the overall function of consciousness is to enable agents internally manipulate their internal representations (of perceptions, thoughts, and actions) supporting goal-directed behaviours. These manipulations produce new knowledge in order to improve the alignment between agent's representations and goals, consequently dealing with unexpected situations and new goals.*

Goal-based Integrated Neural Patterns (GINPs): the neural encoding of conscious contents.

The RIM theory introduces the new concept *Goal-based Integrated Neural Patterns* (GINPs; see Figure 4.5). These are brain active neural representations that (a) are consciously perceived and intentionally manipulable by an agent, and (b) are directly related to the pursued goals. GINPs have a compound nature and each part of them (*sub-GINP*) encodes different aspects leading to goal-directed behaviour (e.g. percepts, motivations, goals, actions).

On the basis of the goal-relevance and the access to consciousness of each sub-GINP, I identify four types of brain representations. *GINP*, conscious representations that has a high level of goal-relevance and stability. *Non-GINPs*, unconscious representations that have weak or no relevance with the goals pursued. *Pre-GINPs*, unconscious representations that have a minor level of goal-relatedness; these representations can influence conscious representations on the basis of unconscious processes (e.g. priming) and do not have the support of top-down attention (see studies on attention and consciousness dissociation, e.g., Koch & Tsuchiya, 2007). *Temp-GINPs*, unstable representations that show a low level of goal-relevance but nevertheless temporary access consciousness (e.g., distractors); although these representations can access consciousness in a transitory way, they lack temporal stability and hence tend to be suppressed.

Brain correlates of GINPs correspond to the activation of a distributed macro-representation involving many structures at multiple levels of the sensory-motor hierarchy (see figure 4.5, right). In particular, sub-GINPs are encoded in different brain macro-systems supporting the goal-directed conscious manipulation of representations.

The four functional components of consciousness.

The RIM theory proposes that consciousness relies on four 'components' (Figure 4.6), supported by partially overlapping anatomo-functional brain macro systems:

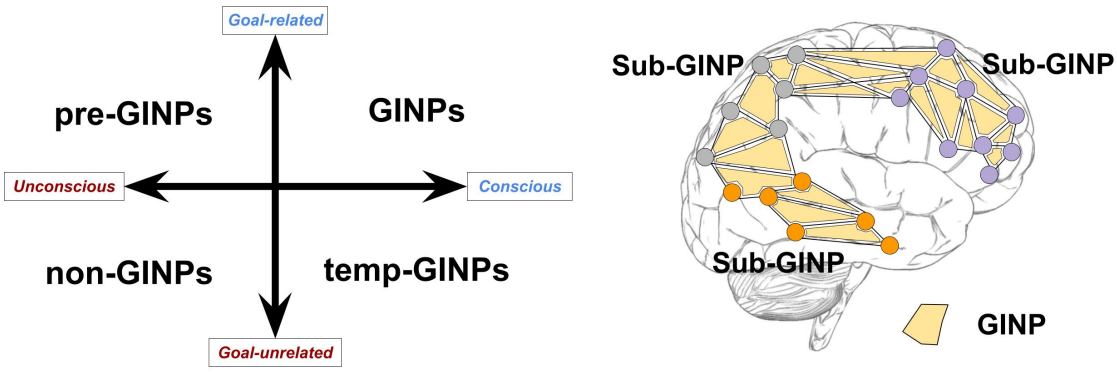


Figure 4.5: On the lefts: different types of GINPs. On the right: image exemplifying the a Goal-based Integrated Neural Pattern (GINP). The whole GINP is composed of four sub-GINPs coloured in orange, grey, and violet, coding for different elements related to a goal (e.g. perceptual features of a goal, affordances related to the goal-achievement, and possible goal-directed action sequences)

(1) Perceptual working memory, (2) Abstract working memory, (3) Internal manipulator, and (4) Motivational component. Following paragraph will describe these components in details.

The *perceptual working memory* is a key component formed by several partially segregated 'unimodal' sub-systems that perform bottom-up sensory processing (Belger et al., 1998; Quak et al., 2015; D'Esposito, 2007). These activations lead to form increasingly abstract percept-related sub-GINPs (e.g., for the visual modality, from low-level visual features - edges and corners - to high-level representations - objects). The same component also supports a top-down information flow through which top-down processes can cause the re-activation of the peripheral sub-GINPs (e.g., imagination processes; Stokes et al., 2009; Zacks, 2008; Kosslyn, 1999). On the one hand, the bottom-up information flows 'propose' pre-GINPs (e.g. percepts) to higher-level cognitive areas. On the other hand, At same time, the top-down manipulation processes favour only the pre-GIMPs that are relevant for the active goal/sub-goals to become conscious. In the brain, the perceptual working memory component is supported by cortical hierarchical pathways, encoding information at multiple levels of abstraction (Felleman & Van Essen, 1991; Mechelli et al., 2004; Baldassarre et al., 2013a). These systems support distributed neural representations corresponding to a perceptual part-GINP. As also proposed by

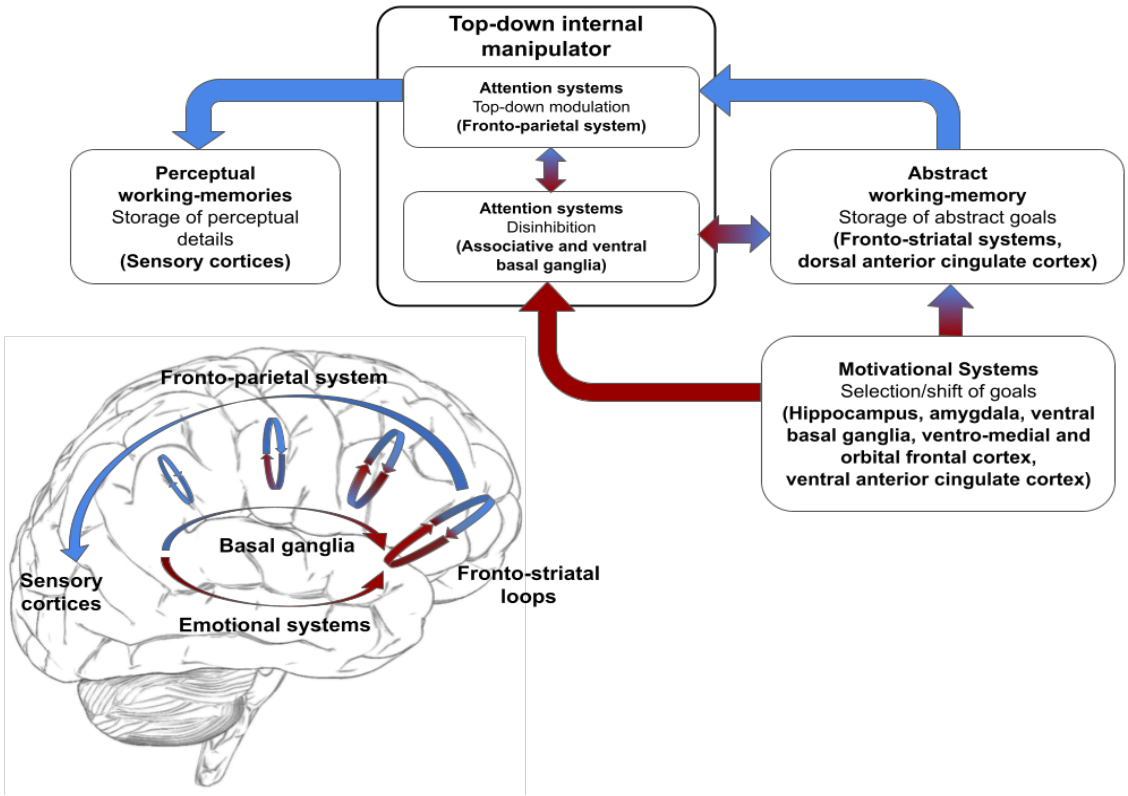


Figure 4.6: Schema showing the 'components' (sets of functionalities) of the RIM theory of consciousness, and their relation with specific anatomo-functional systems of the brain. The red-to-blue coloured gradient indicates the decreasing involvement of emotional/motivational elements, and the 'goal proximity', of the processes implemented in the related brain areas.

the GNWT, fronto-parietal cortical pathways play an important role to support these modal working memories. Moreover, the RIM theory proposes that the basal ganglia-thalamocortical reverberating circuits play a key role to select these representations.

The *abstract working-memory* component plays the key role actively storing and integrating the sub-GINPs related to different aspects of goal-directed processes (e.g. contexts, goals, behavioural strategies, predictions, and values). These sub-GINPs are encoded in a more abstract format with respect to those supported by the perceptual working memory, and so represent a form of meta knowledge. Importantly, sub-GINPs in abstract working memory are functionally linked and encode spatial and dynamical temporal relations of the world elements (environment, objects, and agents). Within the brain, abstract multimodal sub-GINPs are

encoded in representations within different prefrontal cortices (e.g. dlPFC, vlPFC, and ACC; Barraclough et al., 2004; Diamond, 2013) and subcortical areas (e.g. basal ganglia-thalamo-cortical loops; O'Reilly & Frank, 2006). Within each area, neural winner-take-all mechanisms allow the activation of only one or few possible patterns at a time (Aron, 2007). Based on functional links between sub-GINPs in different areas, the component plays a 'hub role' by dynamically integrating its abstract sub-GINPs with the more detailed ones of the perceptual working memory component. This process supports imagination (e.g. visual planning process; Jung et al., 2019), activating a sequence of sub-GINPs encoding the world states to traverse to attain a desired goal.

The key function of the *internal manipulator* component is to select sub-GINPs within the abstract working memory and the perceptual working memory components so as to form a whole GINP having maximum alignment with with the agent's goals/sub-goals. This is a key function as the manipulator acts as the 'attentional scalpel' through which consciousness expresses its adaptive utility. In particular, the manipulator sculpts the active GINP by dynamically adding, removing, and changing the activation of the sub-GINPs that form it. This allows the agent to produce new knowledge to improve the alignment of internal representations with the pursued goal, to accomplish new goals, or accomplish familiar goals in new conditions. Within the brain, the component's operations rely on the disinhibition mechanisms of basal ganglia-thalamo-cortical loops supporting macro selections (Redgrave et al., 1999; Mink, 1996), and on the local inhibitory circuits of the cortical frontal-parietal system performing micro local selections (Fuster & Bressler, 2015; Kappel et al., 2014). The influence of basal ganglia on the cortex has a diminishing gradient moving from frontal to posterior cortical areas; instead, the effect of cortical competitive mechanisms have an increasing gradient moving in the opposite direction.

In the RIM theory the *motivational component* plays the key function of guiding the manipulator to select and activate the sub-GINPs encoding the goals and sub-goals to pursue ('intentions') within the abstract working memory. Moreover,

alongside the active goals/sub-goals, the component contributes to directly drive the internal manipulator to select other sub-GINPs relevant for goal accomplishment within perceptual working memory (e.g. relevant objects or anticipated action outcomes). In the brain, sub-cortical and ventral cortical structures support extrinsic rewards (Panksepp, 1998; Mirolli et al., 2010) while other sub-cortical and dorsal cortical structures support the computation of intrinsic reward signals (Lisman & Grace, 2005; Ribas-Fernandes et al., 2011; Baldassarre, 2011). These systems encode extrinsic motivations, related to biologically or socially salient elements in the environment, and intrinsic motivations, fostering the acquisition of new knowledge from experience. Overall, these processes carry motivational valence originating from the motivational component to the sub-GIMPs of the abstract and perceptual working memories.

The four classes of RIM computational operations.

The integrated functioning of the four components supports the manipulation of internal representations thus improving their alignment with goals. The manipulation relies on four classes of operations called here *RIM operations* (Figure 4.7). They modify the GINP conscious representations and are now considered in detail.

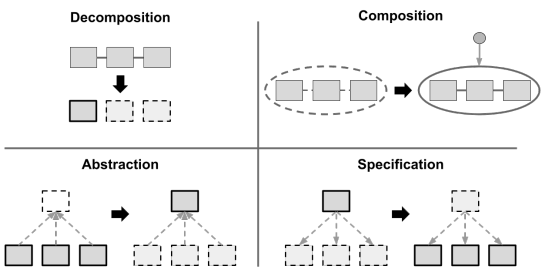


Figure 4.7: The four classes of RIM operations that the manipulator performs on internal representations.

Abstraction leads to the formation of sub-GINPs (e.g. related to world states, goals, and actions) at different levels of abstraction, from the lower-level sub-GINPs (stored in perceptual working memory) to the most abstract ones (stored in abstract working memory). Abstraction allows the manipulation processes to execute a goal-dependent dimensional reduction, capturing in a parsimonious

way goal-relevant aspects of low-level sub-GINPs. Manipulation in particular adapts the level of abstraction depending on the task demands. For example, for the goal of 'taking a tea' a cup might be abstracted in terms of its identity base on its appearing features. Instead, for the goal of 'grasping the cup' it might be represented in terms of shape and position in space while abstracting over colour and texture. In the brain, the processes of abstraction rely on the hierarchically organised stages of cortical pathways which support the goal-directed extraction of increasingly complex features. The acquisition of these representations relies on unsupervised learning processes, possibly affected by reward signals that might bias the acquisition of representations suitable to preserve the information needed to best support goal accomplishment (see the UL/RL integration into perceptual component of the model 3, section 3.3). The basal ganglia-thalamo-cortical macro loops (affective, associative, motor) might facilitate the selection of patterns at different levels of abstraction (Yin & Knowlton, 2006b; Squire et al., 2012).

Specification performs the inverse operations with respect to abstraction, starting from abstract sub-GINPs, for example 'something to drink with', to generate lower-level ones, for example 'my preferred tea cup'. Since specification involves mappings from a few to many features, it requires the goal-based generation of the suitable information at the lower levels depending on the agent's goals (e.g. the imagination of the fine perceptual details of 'my preferred cup'). This process requires the addition of information with respect to the original more abstract representations, requiring a top-down generation of detailed features based on motivational and attentional manipulation processes. I speculate that these generative processes are at the basis of human creativity and productive thinking, allowing the formulation of new solutions to problems. In the brain, specification relies on the top-down 'inverse' activation of cortical pathways, moving from multimodal representations in frontal cortices to modal representations of lower cortices. The generation of the more detailed representations relies also on the cortical and basal-ganglia selection processes biased by motivational/emotional systems and goals.

Decomposition can perform the separation of representations into parts (sub-GINPs) on the basis of motivations and goals. This operation executes an 'horizontal manipulation' at fixed level of abstraction with respect to the 'vertical manipulation' of abstraction and specification. As an example, decomposition could extract the representation of an object (e.g. 'a tea cup') from the background, or the representation of a part of the object (e.g. 'the handle of the cup') from other parts. Decomposition could also select sub-goals at the same level of abstraction, in order to accomplish a final goal. In the brain, decomposition is based on the selection mechanisms involving the cortex and basal ganglia-thalamocortical loops. In particular, it might involve the channels and sub-channels within those loops to disinhibit specific cortical contents, and cortical local winner-take-all mechanisms to facilitate the selection of coherent small neural patterns.

Composition performs the inverse operations with respect decomposition, integrating many sub-GINPs into larger sub-GINPs and a coherent whole GINP. Through composition, the agent can build global items starting from its parts (e.g. to consider a 'cup body', 'handle', 'tea', and 'tea spoon', as a whole 'tea cup'). Composition supports various aspects of goal-directed processes, for example for the creation of plans (e.g. by chunking of a sequence of actions and their effects) or the imagination of the solution of a problem (e.g. the 'candle on the box pinned on the wall'). Note that composition is different from abstraction, in that it does not perform a dimensional reduction (loss of information) but the creation of 'chunks of representations' at the same abstraction level. Despite this, the interaction between composition and abstraction could lead to integrate many sub-GINPs ad to transform them into a more abstract sub-GINP matching the solution to achieve the final goal. In the brain, a chunk can rely on the synchronous activation of the neuronal patterns (e.g. synchronous firing of spiking neurons). Moreover, it can rely on 'horizontal' connections within and between different cortical areas (e.g., encoding two different colours within a visual area, or the 'red' colour encoded in a visual area and the value 'dangerous' in an affective area).

The RIM operations give rise to a super-ordinate function call here *conscious knowl-*

edge transfer (CKT). CKT refers to the capacity of the RIM operations to internally generate the new knowledge that the agent needs to achieve the desired goals. CKT can in particular transfer knowledge from familiar contexts to novel contexts, thus producing the flexibility typical of human cognition and behaviour. To this purpose, CKT operates by flexibly abstracting, specifying, dividing and composing the sub-GINPs encoding the current knowledge (e.g. on objects, goals, actions, and expected outcomes). This allows the agent to build the new knowledge needed to successfully act in novel conditions or accomplish new goals. CKT leads to a self-directed manipulation of the simulated internal reality whose effects in terms of new knowledge production (e.g. new views on situations, elaborations of new plans, problem solving insights) are experienced multi-sensorially and emotionally by the agent. Differently from the concept of *generalisation*, CKT requires the capacity to create knowledge beyond previous experiences. For example, while generalisation involves interpolation processes (e.g. the imagination of a goal involving an object positioned between two previously experienced positions) CKT involves extrapolation processes (e.g. the imagination of an object located anywhere in a known space). This is based on the extraction of relevant regularities from previous experiences (e.g. an isometry, namely the preservation of the spatial ratio during a geometric transformation such as a translation), in order to transfer it to the situation where the new knowledge is needed.

4.2.3 Implications of the RIM framework

The RIM framework has both scientific and technological implications. On one side, it integrates the other theories of consciousness into a computational and coherent framework. This process should prompt the collaborative comparison between the main theories of consciousness, further specifying their computational brain mechanisms. On the other side, it gives indications to build new robotic architectures. A new generation of robots could benefit of a consciousness-like cognitive processing to overcome rigidity issues, developing general intelligence.

Following sections explain in details these implications.

Integrating other theories into a coherent framework.

The RIM theory of consciousness integrates various elements of the main theories of consciousness, building a coherent framework that can benefit of each theory.

The IIT (Tononi, 2008) stresses the importance of a high integration of information into specific computational architectures. This integration is supported by a strong 'cause-effect relationship' between the connections that support these architecture (e.g. those into the cortex-thalamus system). These features make these networks able to support an high level of discriminability between their activation patterns. The RIM theory is strongly linked to the concepts of discriminability and integration in the brain, at the same time enriching the IIT with a functional perspective. First, I expect the perceptual and abstract working memory components to perform a high discrimination of experiences. In particular, the manipulation component selects specific sub-GINP, assigning a specific and stable meaning to the conscious experience. Indeed, a sub-GINP is a representation that shows a strong and stable activation, resulting into an high discriminability with respect non-GINPs activations. Moreover, the selection of GINPs implies high integration. Indeed, GINPs are highly integrated neural patterns encompassing different sub-GINPs located in different areas of brain and encoding goal-related information (e.g. sensory, motor, affective representations). The GINPs thus represent a 'computational glue' allowing the brain to maintain a coherent conscious experience during goal pursuit.

The CDZ theory (Damasio, 1989) ascribes a key role to bottom-up/top-down information flows into the brain sensory hierarchies, involving unimodal cortices (peripheral CDZs) and associative cortices (central CDZs). Moreover, this theory highlights the importance of emotional signals (somatic markers) that influence the computations within the CDZs. The RIM takes into account many concepts of the CDZ theory, further specifying and enriching them with neuroscientific and computational details. First, the RIM theory ascribes a key role to the bidirectional hierarchical brain systems. These activations correspond to the different

parts of the GINPs encoded within the perceptual and abstract working memory components of the RIM system. These bottom-up/top-down flows are supported by the internal manipulator. In particular, it performs abstraction operations, dynamically selecting specific stages into the hierarchy of CDZs, and specification operations, generating new representations into peripheral CDZs. Second, in agreement with the CDZ theory, the RIM ascribes to emotions and motivations the role to assign emotional valence to experience. Indeed, the RIM theory specifies that these processes might take place in terms of the affective systems guiding manipulation operations on sub-GINPs.

The GWT (Baars, 1997) and the GNWT (Dehaene & Changeux, 2011) propose that the activation of a global workspace, supported by high-level brain areas (e.g. the frontal-parietal system), is fundamental to make the relevant information able to access the consciousness. This information is amplified (ignition; Dehaene & Changeux, 2005) and dispatched from the central workpace to many peripheral systems (broadcasting) that can shape the consciousness contents. The RIM theory takes into account many concepts of these theories, enriching them with a further computational specification (manipulation functions) and linking them to the brain mechanisms that lead to a goal-directed behaviour. First, the mechanism of the 'ignitions' is fully in agreement with the mechanism of the GINPs activation proposed by the RIM theory. In particular, the GINPs and sub-GINPs are activated on the basis of a top-down manipulation acting on the central and peripheral systems. This manipulation process could show similar dynamics to that of ignition. Second, the RIM theory ascribes a key role to the frontoparietal brain system, proposing that it is foundamental for the top-down goal-directed control of the sensorimotor cortical pathways. Moreover, while giving a strong importance to the frontal-parietal system, the RIM strongly focus on the interaction between basal ganglia and cortex to select the more suitable sub-GINPs.

The HOTs (Brown et al., 2019) stress the importance of higher-order meta-representations of first-order states. In particular, the Radical Plasticity Theory proposes that the meta-representations show specific features making them able to be consciously

processed, i.e. stability, strength and distinctiveness. Moreover, the HOTs claim that these representations could need of an 'inner awareness' of own cognitive and/or emotional processes (self-consciousness). The RIM theory integrates these concepts with a computational view of the brain. For example, a key feature of the RIM theory is that the manipulation process gives rise to the creation of the sub-GINPs at many stages of sensorymotor hierarchy. This process could correspond to the meta-representational processing suggested by the HOTs. In particular, the RIM assumes that the perceptual representation of a goal emerges due to the influence of affective and attentional process on perceptual working-memory. This goal representations could be considered a meta-representation. The RIM also takes into account the second claim of HOTs regarding the concept of inner awareness. In particular, the internal manipulation of representations, which can also involve the agent itself, could be equated to, and specifies, the inner awareness of HOTs.

The sensorimotor theory (O'Regan & Noe, 2001) emphasise the relevance of the sensorimotor interactions that conscious agents engage with the environment. In particular, it supports the idea that the consciousness requires the *alignment* between the agent's internal processes (e.g. perceptual representations) and the external world. The RIM integrates these ideas into an embodied and computational view. In particular, it proposes that consciousness plays a relevant role in the adaptation of agents during their interaction with the environments. In accord with this view, the RIM proposes that consciousness facilitates the production of more adaptive behaviours thanks to its internal manipulation, in turn leading to the enhancement of goal-directed processes.

The RIM starts to integrate the key concepts proposed by other theories of consciousness, instantiating a collaborative comparison and creating a coherent functional framework. In particular, it successfully accounts for the concepts Integration/discriminability (IIT), Hierarchical sensory-motor brain organisation (CDZT), top-down control linked to a whole brain information amplification (GWT/GNWT), meta-representational process and metacognition (HOTs) and

sensorymotor agent-environment interactions (SMT). Future theoretical and computational developments will built on the RIM to overcome the present issue reported by investigations on consciousness (e.g. the experimental support of the major theories of consciousness; Yaron et al., 2022; Del Pin et al., 2021; Doerig et al., 2021; Melloni et al., 2021).

Building a new generation of consciousness-inspired robots

Cognitive Robotics and Neuro-robotics explicitly attempt to create synergies between cognitive neuroscience, AI and robotics, with both a scientific aim (investigation of cognition and brain) and a technological aim (building more efficient robots). Machine consciousness (MC) partially overlap with these fields, aiming to define the key elements that an AI and robotic system should have to show a certain level of consciousness (Aleksander, 1995; Gamez, 2008; Reggia, 2013). In agreement with these fields, here I assume that some functional and architectural elements of consciousness can be introduced in robotic systems to enhance them. I now show some application cases, showing that the RIM can boost AI and robotics fields.

Flexibility. A widely recognised important limitation of current AI systems is their limited flexibility, intended as the difficulty to face new tasks and/or novel conditions, and to reason with incomplete information (Hassabis et al., 2017; Lake et al., 2017; Marcus & Davis, 2019). The central idea of the RIM theory is that the flexibility exhibited by the brain depends on the capacity to internally manipulate the representations of goal-directed elements (objects, goals, actions, etc.) so as to adapt them to pursue new goals in possibly new conditions. This could provide robotic agents the capacity to actively re-adapt and complete the knowledge acquired in previous experiences to face novel goals and conditions (conscious knowledge transfer).

Learning speed. A second important limitation ascribed to current AI systems is their need for much data and training time to acquire the capacity to solve tasks (Lake et al., 2017; Marcus & Davis, 2019; Ullman, 2019). Conscious knowledge transfer could be a relevant mechanism used to speed up learning as it would represent a powerful way to transfer knowledge between tasks and domains, operating alongside the standard generalisation capabilities of neural networks. Moreover, within RIM architectures intrinsic motivations could guide not only the focused search of the knowledge that the agent needs in the environment, as it is commonly done, but they could also guide the internal construction of knowledge, as further discussed below.

Creativity. Creativity and imagination are other cognitive capacities that are strongly limited in AI systems (Hassabis et al., 2017; Lake et al., 2017; Marcus & Davis, 2019). As also observed by these authors, generativity offers a solution to this problem. The RIM theory proposes that generativity, integrated with manipulation, could in particular be used to modify the goal-directed elements to build new solutions. Instead, such elements are usually considered fixed in AI systems.

Human-friendly AI. Many authors advocate an AI capable of usefully interacting with humans and of aligning with their values (Harari, 2016; Bostrom, 2014). AI architectures encompassing some elements of consciousness suggested by the RIM theory should facilitate this for a number of reasons. First, as discussed above, these architectures could exhibit a higher degree of flexibility, and this represents an important quality to more easily interact with humans. Second, they would have a richer capacity to reason about affective issues, an important element to support a suitable interaction with humans (Huang et al., 2019). Last, they would have a sophisticated motivation component that could facilitate the design of value systems closer to those of humans (Dignum, 2018).

Open challenges. Machine consciousness indicate the main elements that are essential for having conscious intelligent robotics systems. In particular, it proposes the five axiomatic elements of artificial consciousness (world models, imagination, attention, planning, and affective evaluation; Aleksander, 1995) and the main approaches to model conscious cognition (self-modelling, information broadcasting, higher-level representations, attention processes, and information integration; Reggia, 2013). Overall, a RIM-based robotic architecture would include all these elements. However, it could still be missing subtle elements.

First, the integration of those elements necessary for consciousness is a fundamental open challenge. Much of the brain flexibility might indeed rest on its architecture, evolved in millions of years by the huge-scale 'genetic algorithm' of evolution, integrating all habitual and goal-directed processes analysed here in a harmonic way (Baldassarre et al., 2017; Ullman, 2019). Second, beyond macro-functions proposed by RIM, the brain could exploit 'lower-level functionalities' that might be very important to have consciousness. For example, the brain has an high capacity for creating associations that rests on its grid-like circuits (vs. present artificial neural-network architectures privilege bottom-up/top-down directional information flows; Lynn & Bassett, 2019). The brain shows an highly dynamic nature based on attractors, that might be fundamental for the effective functioning of the RIM mechanisms operating on GINPs (Breakspear, 2017).

Future developments of RIM aim to further clarify these key feature to achieve a human-like artificial consciousness.

Chapter 5

General conclusions

The following sections explain in details the achievements and future directions of this project, passing trough the three main project phases anticipated in Introduction (Figure 1.1, i.e. 'Main', 'Application', 'Post-doc').

5.1 Summary of achievements and contributions to knowledge of this project

This PhD research project mainly focuses on the investigation of neurocognitive processes that allow humans to express cognitive and behavioural flexibility. In particular, I have focused on the capacity of humans to influence, under the guidance of goals, own representations, in case of both healthy and pathological conditions.

The first step of this project has investigated the scientific literature regarding the key concepts for the main topics, such as goal-directed behaviour, executive functions, internal representations and perceptual categorisation (section 2.1). The literature review has highlighted that flexible and adaptive behaviours (e.g. categorisation or planning) require both the creation and selection of useful perceptual representations. For example, the emergence of category-based representations (categorical perception) is influenced by task-dependent and task-independent

signals while top-down attention processes perform a task-directed modulation of perceptual representations to support adaptive behaviours. Moreover, the literature review has highlighted that inner speech, a self-directed form of language, is able to influence the previous high-order processes.

Furthermore, I have investigated the computational approaches regarding the previous topics (section 2.2), in that the computational literature shows a landscape of formal models systematising many experimental evidences. This review has highlighted that both models of categorical perception and cognitive flexibility clarify many neuro-cognitive processes but they neglect other key ones. For example, no model of categorical perception proposes an investigation of the interaction between motivational, perceptual and motor brain systems. In particular, no model emulates the interaction of associative and reinforcement learning mechanisms underpinning these systems interactions. Regarding the models of flexible cognition, they totally neglect a representational aspect of cognitive flexibility, i.e. the top-down selection of suitable perceptual representations. At last, there are few models of inner speech and none investigate its contribution to the solution of a neuropsychological task requiring flexible cognition.

On the basis of these analysis, I have proposed a formal theory of flexible cognition defined 'three-components hypothesis' (section 2.3). This theory formalises the interaction between the brain systems allowing humans to execute a flexible behaviour on the basis of a goal-directed manipulation of perceptual representations. An extension of the three-components hypothesis (section 2.3) has included the role of inner speech as a second-order manipulation, suggesting that humans benefit of both a perceptual and a conceptual self-directed manipulation of representations to express a flexible behaviour. Overall, the three-components hypothesis remarks the key idea that humans are able to adapt both the surrounding world and, above all, themselves.

The three-components framework focuses on a top-down selection of representations but it does not investigate the representations learning/acquisition processes. The theory in fact expects that, before the development of the ability to

goal-directly manipulate the one's own representations, repeated sensory-motor interactions lead infants and children to acquire adaptive perceptual representations. To model these learning aspects, I have proposed 'the motivated categorical perception hypothesis' (section 2.4), formalising the relationship between motivational, perceptual and motor system leading to the emergence of a categorical perception.

On the basis of my theoretical proposals I have built three novel computational models (chapter 3). The validation of these models against human experimental data (1) corroborates the goodness of theoretical premises and (2) conveys scientific insights regarding human cognition.

The first model (section 3.1) corroborates the three-components hypothesis, for which the interaction of sensory-motor loops and a top-down goal-directed manipulation of perceptual representations describes the emergence of humans flexible cognition and behaviour. The model explains many experimental data obtained during the performance of WCST presented in already validated and published works, from healthy young and old adults and pathological populations (patients with frontal lesions and Parkinson patients).

This model is the first that emulates the top-down manipulation of perceptual representations during the performance of WCST (see section 3.1.5 for a computational review of the other models) but it neglects the self-directed manipulation of high-order representations (e.g. working memory). Moreover mostly computational models of language neglect its role in cognitive control and attention (see section 3.2.4 for a computational review of the other models).

Overcoming these limits, I have proposed an update of the first model (section 3.2), showing the addition of an inner-speech component as a second-order kind of self-directed manipulation. The model explains humans experimental data, extracted from already validated and published works, obtained from an experimental protocol integrating the WCST and a verbal shadowing protocol (i.e. an experimental protocol that disrupts the contribution of inner speech). In par-

ticular, it disentangles and clarifies the role of inner-speech as a second-order self-directed manipulation. More generally, the model corroborates the extended three-component hypothesis, for which self-directed forms of both perceptual and high-order manipulations support the expression of flexible cognition and behaviour.

The three-components hypothesis expects that early sensory-motor learning processes lead to the emergence of suitable representations (e.g., categorical perception), which in turn are targeted by top-down manipulation processes. Interestingly, despite the models of categorical perception clarify some aspects of these learning processes, they neglects the investigation of different underpinning learning mechanisms (i.e. associative and reward-based learning) involving both the motivational, perceptual and motor systems.

The third model (section 3.3) overcomes these limitations operationalising the motivated categorical perception hypothesis, for which the interaction of perceptual/motivational/motor systems and the subsequent balance between associative and reward-based mechanisms lead to the acquisition of adaptive perceptual representations. The model results show that different perceptual learning profiles could explain two opposing views regarding the categorisation skills in autistic people. In particular, the model suggests that there are different perceptual profiles in autism and that both an extreme reward-dependent and an extreme reward-independent learning profiles could explain the divergent perceptual skills in autistic people.

The major aim of the previous models is the corroboration of the theoretical premises, passing a validation that considers the experimental data obtained from human healthy and clinical populations. However, this validation process has also produced many scientific insights regarding human cognition and behaviour. For example, these investigations have had also clinical implications, clarifying the link between the top-down manipulation mechanisms, the categorical perception, and specific subsequent visible behavioural deficits. On the other hand both the theoretical frameworks and models represent important tools for the development

of further traslational researches (chapter 4).

For example section 4.1 shows that the three-components hypothesis and the model 2 represent a tool in computational psychiatry. In particular, the model has reproduced the WCST experimental data, extracted from already published works, obtained from neurotypical and autistic populations of different ages (children, teenagers, young adults, middle adults). The results suggest that the inner-speech acquires a growing supporting role along the life span of neurotypical people while it is extremely reduced in autistic individuals. This conclusion highlights that a clinical treatment that focus on an early development of inner speech in autistic children could represent a developmental support for autistic people along the life span.

At last the theories and models developed in this project have guided the specification of many theoretical and neuroscientific aspects of consciousness (section 4.2) with also technological implications (section 4.2.3). In particular, I have integrated the three-components hypothesis and many element of other theories of consciousness, proposing the four-component theory or 'Representations Internal Manipulation (RIM) theory of consciousness'. This theoretical framework proposes a unifying theory of consciousness and describes the brain processes at the basis of the conscious and flexible manipulation of internal representations. Moreover, the RIM framework can lead to the development of a new generation of consciousness-inspired robotics architectures (for a review on machine consciousness see Reggia, 2013) that could overcome the current limits of robotics, that is, achieving a human-level flexibility and creativity.

Table 5.1 and table 5.2 summarise the main features and achievements of this project. In particular, the first shows the main achievements while the second shows the application cases of my studies. The tables clearly show that each of the three theories of cognition leads to the development of one computational model, that explains human experimental data and provides insights regarding human cognition.

Process investigated	Theoretical proposal	Computational model	Performed task	Human populations 'fitted'
Cognitive flexibility, Representation manipulation	Three-components hypothesis	Model 1	WCST	Healthy young adults Healthy old adults Frontal patients Parkinson patients
Cognitive flexibility, Representation manipulation, **and inner-speech**	Three-components hypothesis **(extended with language)**	Model 2	WCST (verbal shadowing)	Young adults: control condition motor tapping verbal shadowing
Categorical perception, Motivational systems, Representation learning	Motivated categorical perception hypothesis	Model 3	Categorisation task	ASC variability

Table 5.1: The table reports the projects achievements. In particular, it shows that each scientific topic is formalised (theoretical proposal) and guides the building of a computational model. Then the computational model is tested and the results are compared with those of many human populations. WCST: Wisconsin Cards Sorting test. ASC: Autism Spectrum Conditions

Application case	Theoretical framework	Model	Achievements
Clinical sciences	Three-components hypothesis **(extended)**	Model 2	Scientific insights (Inner speech and autism)
Neuro-cognitive Science Robotics	Consciousness Executive functions Goal-directed behaviour	Four-components hypothesis (RIM)	Unifying theory of consciousness Robotics insights

Table 5.2: The table describes two application cases of my theories and models. RIM: Representations Internal Manipulation theory.

5.2 Limitationss and future directions

The theories and the models presented in this project show limitations that will be starter points for further scientific investigations.

For example, my frameworks describe the processes supporting human flexible cognition and the emergence of a categorical perception. However, although I have already proposed that they could represent opposite poles of a continuum in a evolutionary prospective (i.e. from an emergent categorical perception to a goal-directed manipulation), I have not developed an integrated framework yet. A new framework could explain these relationships along the human cognitive development.

Moreover, the three-components hypothesis describes the human neuro-cognitive processes supporting the cognitive flexibility but it does not consider the sequential aspects of higher executive functions such as planning and problem solving. The four-components theory or RIM starts to describe the functioning of these

processes along an extended time windows (planning), but it still represents a preliminary theoretical work that will be further developed.

As suggested by specific sections of this project, further research will involve many computational updates of the models presented in this work. For example in section 3.1.5 I explained that the model 1 and 2 could benefit of an update of the working memory component, replaceable with a distributed recurrent network (e.g. echostates networks; Mannella & Baldassarre, 2015b). This update could reproduce the gating mechanisms of working memory in a more bio-grounded way, making the model WM more adapt to support different tasks. Again, these models could benefit of an update of the internal attention mechanisms (manipulator) and the external attention mechanism (retina displacement). The first update should make them able to focus on different perception stages (i.e. both global objects and specific features) and hence to investigate the human global/local perceptual interactions in case of clinical conditions (e.g. autism; Koldewyn et al., 2013). The second update should support a bio-inspired bottom-up retina movement (Baldassarre et al., 2019b) and hence the investigation of the bottom-up/top-down interaction supporting the overt attention. At last, the section 3.3.5 suggests many technical improvements that the third model could receive. For example the model is not able to face the catastrophic forgetting and adaptive categorical perception. The integration of this model with mechanisms supporting the first and the second model (e.g. a dynamical internal manipulator) could overcome this issue, also allowing the investigation of a balance between perceptual learning and internal attention. At last, an update of the learning rules of the model layers could support further investigations regarding the perceptual and motivational interactions during the emergence of categorical perception.

Beyond these methodological improvements, I aim to implement a further computational model that emulates the interaction of planning processes and the goal-directed representations manipulation. Starting from the four-component/RIM theory, this new model could clarify the role that the internal representations manipulation acquired for different executive functions. In particular, the compu-

tational models of human planning processes (Stewart & Eliasmith, 2011; Zarr & Brown, 2019; Bieszczad & Kuchar, 2015; Donnarumma et al., 2016) tend to focus on the actions and the external states of the world while this new model would demonstrate that the 'internal actions' (representations manipulations) have a key role during the creation of an efficient plan.

Overall, the project suggests many promising scientific directions regarding the clinical applications. For example, my studies highlight the possibility to extend the investigations of goal-directed manipulation/learning in computational psychiatry. First, future works aim to test the models 1 and 2 against further data obtained from psychiatric patients that show perceptual deficit and/or inner-speech alteration (e.g. schizophrenic and PSTD patients). This process could provide many insights regarding the alteration of perceptual and executive processes in many psychiatric conditions. With few adaptations, also the model 3 could be used to fit and explain specific experimental data (e.g. behavioural indices, as done in the 1/2 models) from different population of psychiatric patients. In particular, it could be used to disentangle the different contributions of motor and perceptual skills during the expression of a goal-directed behaviour of autistic children in a motor sorting task (Taffoni et al., 2019).

At last, the algorithms exploited in this project will guide a deepening in neurorobotics. In particular, as suggested in section 4.2.3, robotics architectures could benefit of a goal-directed internal manipulation of representations to support open-ended learning mechanisms in unpredictable environments. This improvement could make robots more autonomous and flexible in case of low human interventions, and more safe in case of human-machine interactions.

Bibliography

Adams, C., Lockton, E., Freed, J., Gaile, J., Earl, G., McBean, K., Nash, M., Green, J., Vail, A., & Law, J. (2012). The social communication intervention project: a randomized controlled trial of the effectiveness of speech and language therapy for school-age children who have pragmatic and social communication problems with or without autism spectrum disorder. *International Journal of Language & Communication Disorders*, *47*(3), 233–244.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, *8*(10), 457–464.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, *141*(5), 931.

Aleksander, I. (1995). Artificial neuroconsciousness an update. In *International Workshop on Artificial Neural Networks*, (pp. 566–583). Springer.

Aman, M. G. (2005). Treatment planning for patients with autism spectrum disorders. *Journal of clinical psychiatry*, *66*, 38.

Amari, S.-I. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5*(4-5), 185–196.

Ambery, F. Z., Russell, A. J., Perry, K., Morris, R., & Murphy, D. G. (2006). Neuropsychological functioning in adults with asperger syndrome. *Autism*, *10*(6), 551–564.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of cognitive neuroscience*, *12*(3), 505–519.

Arbib, M. A. (2008). From grasp to language: Embodied concepts and the challenge of abstraction. *Journal of Physiology-Paris*, *102*(1-3), 4–20.

Aron, A. R. (2007). The neural basis of inhibition in cognitive control. *The neuroscientist*, *13*(3), 214–228.

Arsenault, J. T., Nelissen, K., Jarraya, B., & Vanduffel, W. (2013). Dopaminergic reward signals selectively decrease fmri activity in primate visual cortex. *Neuron*, *77*(6), 1174–1186.

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annul Review of Psychology*, *56*, 149–178.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147.

Association, A. P., Association, A. P., et al. (2013). Dsm 5. *American Psychiatric Association*, *70*.

Astafiev, S. V., Stanley, C. M., Shulman, G. L., & Corbetta, M. (2004). Extrastriate body area in human occipital cortex responds to the performance of motor actions. *Nature neuroscience*, *7*(5), 542–548.

Baars, B. J. (1997). In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, *4*(4), 292–309.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45–53.

Baars, B. J., Franklin, S., & Ramsoy, T. Z. (2013). Global workspace dynamics: cortical "binding and propagation" enables conscious contents. *Frontiers in psychology*, *4*.

Baars, B. J., Ramsøy, T. Z., & Laureys, S. (2003). Brain, conscious experience and the observing self. *Trends in neurosciences*, *26*(12), 671–675.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baldassarre, G. (2011). What are intrinsic motivations? a biological perspective. In A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P.-Y. Oudeyer, M. Schlesinger, & Y. Nagai (Eds.) *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, (pp. e1–8). Frankfurt Institute of Advanced Studies (FIAS), New York, NY: IEEE. Frankfurt am Main, Germany, 24–27/08/11.

Baldassarre, G., Caligiore, D., & Mannella, F. (2013a). The hierarchical organisation of cortical and basal-ganglia systems: a computationally-informed review and integrated hypothesis. In G. Baldassarre, & M. Mirolli (Eds.) *Computational and Robotic Models of the Hierarchical Organisation of Behaviour*, (pp. 237–270). Berlin: Springer-Verlag.

Baldassarre, G., Lord, W., Granato, G., & Santucci, V. G. (2019a). An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. *Frontiers in Neurorobotics*, *13*(45).

Baldassarre, G., Lord, W., Granato, G., & Santucci, V. G. (2019b). An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. *Frontiers in neurorobotics*, *13*, 45.

Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K., & Mirolli, M. (2013b). Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks*, *41*, 168–187.

Baldassarre, G., & Mirolli, M. (Eds.) (2013). *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer.

Baldassarre, G., Santucci, V. G., Cartoni, E., & Caligiore, D. (2017). The architecture challenge: Future artificial-intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction. *Behavioral and Brain Sciences*, *40*(40), e254.

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science*, *344*(6182), 424–427.

Baldo, J. V., Dronkers, N. F., Wilkins, D., Ludy, C., Raskin, P., & Kim, J. (2005). Is problem solving dependent on language? *Brain and language*, *92*(3), 240–250.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407–419.

Barceló, F. (1999). Electrophysiological evidence of two different types of error in the wisconsin card sorting test. *Neuroreport*, *10*(6), 1299–1303.

Barceló, F., & Knight, R. T. (2002). Both random and perseverative errors underlie wcst deficits in prefrontal patients. *Neuropsychologia*, *40*(3), 349–356.

Barceló, F., Suwazono, S., & Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature neuroscience*, *3*(4), 399–403.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, *80*, 169–189.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature neuroscience*, *7*(4), 404.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davids, & D. G. Beiser (Eds.) *Models of Information Processing in the Basal Ganglia*, (pp. 215–232). Cambridge, MA: The MIT Press.

Barto, A. G., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology – Cognitive Science*, *4*(907), e1–15. Edited by: Tom Stafford, University of Sheffield, UK Reviewed by: Karl Friston, University College London, UK; Nathan F.Lepora, The University of Sheffield, UK.

Basanisi, R., Brovelli, A., Cartoni, E., & Baldassarre, G. (2020). A generative spiking neural-network model of goal-directed behaviour and one-step planning. *PLoS Computational Biology*, *16*(12), e1007579.

Bauer, A. J., & Just, M. A. (2017). A brain-based account of "basic-level" concepts. *Neuroimage*, *161*, 196–205.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive behavior*, *11*(4), 209–243.

Belger, A., Puce, A., Krystal, J. H., Gore, J. C., Goldman-Rakic, P., & McCarthy, G. (1998). Dissociation of mnemonic and perceptual processes during spatial and nonspatial working memory using fmri. *Human brain mapping*, *6*(1), 14–32.

Bengio, E., Thomas, V., Pineau, J., Precup, D., & Bengio, Y. (2017). Independently controllable features. *ArXiv*, *abs/1703.07718*.

Bengio, Y. (2017). The consciousness prior. *arXiv preprint arXiv:1709.08568*.

Berdia, S., & Metz, J. (1998). An artificial neural network stimulating performance of normal subjects and schizophrenics on the wisconsin card sorting test. *Artificial intelligence in medicine*, *13*(1-2), 123–138.

Bieszczad, A., & Kuchar, S. (2015). Neurosolver learning to solve towers of hanoi puzzles. In *2015 7th International Joint Conference on Computational Intelligence (IJCCI)*, vol. 3, (pp. 28–38). IEEE.

Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., & Busemeyer, J. R. (2010). Sequential learning models for the wisconsin card sort task: Assessing processes in substance dependent individuals. *Journal of mathematical psychology*, *54*(1), 5–13.

Böhmer, W., Springenberg, J. T., Boedecker, J., Riedmiller, M., & Obermayer, K. (2015). Autonomous learning of state representations for control: An emerging field aims to autonomously learn state representations for reinforcement learning agents from their real-world sensor observations. *KI-Künstliche Intelligenz*, *29*(4), 353–362.

Bókkon, I., Salari, V., Scholkmann, F., Dai, J., & Grass, F. (2013). Interdisciplinary implications on autism, savantism, asperger syndrome and the biophysical picture representation: Thinking in pictures. *Cognitive Systems Research*, *22*, 67–77.

Bonnasse-Gahot, L., & Nadal, J.-P. (2020). Categorical perception: A groundwork for deep learning. *arXiv preprint arXiv:2012.05549*.

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Flexibility, situatedness, language and sociality in abstract concepts reply to comments on "words as social tools: Language, sociality and inner grounding in abstract concepts". *Physics of Life Reviews*, *7*, 8.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience*, *35*(25), 9329–9335.

Bracci, S., Ritchie, J. B., & de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, *105*, 153–164.

Braver, T. S., & Bongiolatti, S. R. (2002). The role of frontopolar cortex in subgoal processing during working memory. *Neuroimage*, *15*(3), 523–536.

Breakspear, M. (2017). Dynamic models of large-scale brain activity. *Nature Neuroscience*, *20*(3), 340–352.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in cognitive sciences*, *23*(9), 754–768.

Brunel, N., & Wang, X.-J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of computational neuroscience*, *11*(1), 63–85.

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron*, *76*(4), 838–846.

Bush, G., Vogt, B. A., Holmes, J., Dale, A. M., Greve, D., Jenike, M. A., & Rosen, B. R. (2002). Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proceedings of the National Academy of Sciences*, *99*(1), 523–528.

Cabeza, R., & Dennis, N. A. (2012). Frontal lobes and aging: deterioration and compensation. *Principles of frontal lobe function*, *2*, 628–652.

Caligiore, D., Arbib, M. A., Miall, C. R., & Baldassarre, G. (2019a). The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. *Neuroscience and Biobehavioral Reviews*, *100*, 19–34.

Caligiore, D., Arbib, M. A., Miall, R. C., & Baldassarre, G. (2019b). The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. *Neuroscience & Biobehavioral Reviews*, *100*, 19–34.

Caligiore, D., Borghi, A., Parisi, D., & Baldassarre, G. (2010). Tropicals: A computational embodied neuroscience model of compatibility effects. *Psychological Review*, *117*(4), 1188–1228.

Cangelosi, A., Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Connection Science*, *12*(2), 143–162.

Canini, K. R., Shashkov, M. M., & Griffiths, T. L. (2010). Modeling transfer learning in human categorization with the hierarchical dirichlet process. In *ICML*.

Caras, M. L., & Sanes, D. H. (2017). Top-down modulation of sensory cortex gates perceptual learning. *Proceedings of the National Academy of Sciences*, *114*(37), 9972–9977.

Carcani-Rathwell, I., Rabe-Hasketh, S., & Santosh, P. J. (2006). Repetitive and stereotyped behaviours in pervasive developmental disorders. *Journal of Child Psychology and Psychiatry*, *47*(6), 573–581.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, *37*(1), 54–115.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Carvalho, P. F., & Goldstone, R. L. (2016). *Human Perceptual Learning and Categorization*, chap. 10, (pp. 223–248). John Wiley and Sons, Ltd.

Casasanto, D. (2008). Who's afraid of the big bad whorf? crosslinguistic differences in temporal language and thought. *Language learning*, *58*, 63–79.

Casey, M. C., & Sowden, P. T. (2012). Modeling learned categorical perception in human vision. *Neural networks*, *33*, 114–126.

Caso, A., & Cooper, R. P. (2017). A model of cognitive control in the wisconsin card sorting test: Integrating schema theory and basal ganglia function. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*, (pp. 210–215).

Caso, A., & Cooper, R. P. (2020). A neurally plausible schema-theoretic approach to modelling cognitive dysfunction and neurophysiological markers in parkinson's disease. *Neuropsychologia*, *140*, 107359.

Cella, M., Bishara, A. J., Medin, E., Swan, S., Reeder, C., & Wykes, T. (2014). Identifying cognitive remediation change through computational modelling—effects on reinforcement learning in schizophrenia. *Schizophrenia bulletin*, *40*(6), 1422–1432.

Cepeda, N. J., Kramer, A. F., & Gonzalez de Sather, J. (2001). Changes in executive control across the life span: examination of task-switching performance. *Developmental psychology*, *37*(5), 715.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, *2*(3), 200–219.

Chelazzi, L., Perlato, A., Santandrea, E., & Della Libera, C. (2013). Rewards teach visual selective attention. *Vision research*, *85*, 58–72.

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., Dong, H.-M., Yang, Z., Zang, Y.-F., Zuo, X.-N., et al. (2015). Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. *PLoS One*, *10*(12), e0144963.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, (pp. 1597–1607). PMLR.

Chiu, E.-C., Wu, W.-C., Hung, J.-W., & Tseng, Y.-H. (2018). Validity of the wisconsin card sorting test in patients with stroke. *Disability and rehabilitation*, *40*(16), 1967–1971.

Christakou, A., Murphy, C., Chantiluke, K., Cubillo, A., Smith, A., Giampietro, V., Daly, E., Ecker, C., Robertson, D., Murphy, D., et al. (2013). Disorder-specific functional abnormalities during sustained attention in youth with attention deficit hyperactivity disorder (adhd) and with autism. *Molecular psychiatry*, *18*(2), 236–244.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: The MIT Press.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annu Rev Neurosci*, *33*, 269–298.

Clark, A. (1998). Magic words: How language augments human computation. *Language and thought: Interdisciplinary themes*, (pp. 162–183).

Cleeremans, A. (2011). The radical plasticity thesis: how the brain learns to be conscious. *Frontiers in psychology*, *2*, 86.

Clifford, A., Franklin, A., Davies, I. R., & Holmes, A. (2009). Electrophysiological markers of categorical perception of color in 7-month old infants. *Brain and cognition*, *71*(2), 165–172.

Collins, J. A., & Olson, I. R. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic bulletin & review*, *21*(4), 843–860.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci*, *3*(3), 201–215.
URL `http://dx.doi.org/10.1038/nrn755`

Cox, M., Alavi, Z., Dannenhauer, D., Eyorokon, V., Munoz-Avila, H., & Perlis, D. (2016). Midca: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30.

Da Rold, F. (2018). Defining embodied cognition: The problem of situatedness. *New Ideas in Psychology*, *51*, 9–14.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, *1*(1), 123–132.

Damper, R. I., & Harnad, S. (2000). Neural network models of categorical perception. *Perception & psychophysics*, *62*(4), 843–867.

Daselaar, S., Cabeza, R., Ochsne, K., & Kosslyn, S. (2013). Age-related decline in working memory and episodic memory: Contributions of the prefrontal cortex and medial temporal lobes. *The Oxford handbook of cognitive neuroscience*, *1*, 456–472.

Dasgupta, S., & Osogami, T. (2016). Theory of stdp in dynamic boltzmann machines for learning temporal patterns. *Advances in Neuroinformatics IV*, (p. 62).

Davis, B. L., & MacNeilage, P. F. (2000). An embodiment perspective on the acquisition of speech perception. *Phonetica*, *57*(2-4), 229–241.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge, MA: The MIT Press.

De Baene, W., Ons, B., Wagemans, J., & Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learning & Memory*, *15*(9), 717–727.

de Beeck, H. P. O., Baker, C. I., DiCarlo, J. J., & Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *Journal of Neuroscience*, *26*(50), 13025–13036.

de Zilva, D., & Mitchell, C. J. (2012). Effects of exposure on discrimination of similar stimuli and on memory for their unique and common features. *Quarterly Journal of Experimental Psychology*, *65*(6), 1123–1138.

Dehaene, S., & Changeux, J.-P. (1991). The wisconsin card sorting test: Theoretical analysis and modeling in a neuronal network. *Cerebral cortex*, *1*(1), 62–79.

Dehaene, S., & Changeux, J.-P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *3*, e141.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998a). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 14529–14534.

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998b). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, *95*(24), 14529–14534.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1), 1–37.

Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M., & Wierzchoń, M. (2021). Comparing theories of consciousness: why it matters and how to do it. *Neuroscience of Consciousness*, *2021*(2), niab019.

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755).

Dennis, N., & Cabeza, R. (2012). Frontal lobes and aging: deterioration and compensation. *Principles of frontal lobe function*, *2*, 628–652.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 761–772.

DeYoe, E. A., Carman, G. J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., & Neitz, J. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences*, *93*(6), 2382–2386.

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135–168.

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, *20*, 1–3.

Dijkstra, N., Zeidman, P., Ondobaka, S., Gerven, M., & Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific reports*, *7*(1), 5677.

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive neuroscience*, *12*(2), 41–62.

Donnarumma, F., Maisto, D., & Pezzulo, G. (2016). Problem solving as probabilistic inference with subgoaling: explaining human successes and pitfalls in the tower of hanoi. *PLoS computational biology*, *12*(4), e1004864.

Dove, G. (2018). Language as a disruptive technology: abstract concepts, embodiment and the flexible mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170135.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7-8), 961–974.

Duncker, K. (1935). *Zur Psychologie des produktiven Denkens [The psychology of productive thinking]*. Berlin: Springer.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *science*, *338*(6111), 1202–1205.

Emadi, N., & Esteky, H. (2014). Behavioral demand modulates object category representation in the inferior temporal cortex. *Journal of neurophysiology*, *112*(10), 2628–2637.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in cognitive sciences*, *15*(2), 85–93.

Fagg, A. H., & Arbib, M. A. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw*, *11*(7-8), 1277–1303.

Fallon, S. J., Mattiesing, R. M., Muhammed, K., Manohar, S., & Husain, M. (2017). Fractionating the neurocognitive mechanisms underlying working memory: independent effects of dopamine and parkinson's disease. *Cerebral Cortex*, *27*(12), 5727–5738.

Farreny, A., del Rey-Mejías, Á., Escartin, G., Usall, J., Tous, N., Haro, J. M., & Ochoa, S. (2016). Study of positive and negative feedback sensitivity in psychosis using the wisconsin card sorting test. *Comprehensive Psychiatry*, *68*, 119–128.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, *1*(1), 1–47.

Ferdinando, A. D., & Parisi, D. (2004). Internal representations of sensory input reflect the motor output with which organisms respond to the input. In *Seeing, thinking and knowing*, (pp. 115–141). Springer.

Fernandes, F. D. M., Amato, C. A. D. L. H., & Molini-Avejonas, D. R. (2012). Language therapy results with children of the autism spectrum. *Revista de Logopedia, Foniatría y Audiología*, *32*(1), 2–6.

Figueroa, I. J., & Youmans, R. J. (2013). Failure to maintain set: a measure of distractibility or cognitive flexibility? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, (pp. 828–832).

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, *18*(11), 1664–1671.

Fiore, V. G., Sperati, V., Mannella, F., Mirolli, M., Gurney, K., Firston, K., Dolan, R. J., & Baldassarre, G. (2014). Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot. *Frontiers in Psychology*, *5*(124), e1–17.

Fisk, J. E., & Sharp, C. A. (2004). Age-related impairment in executive functioning: Updating, inhibition, shifting, and access. *Journal of clinical and experimental neuropsychology*, *26*(7), 874–890.

Flippin, M., & Hahs-Vaughn, D. L. (2020). Parent couples' participation in speech-language therapy for school-age children with autism spectrum disorder in the united states. *Autism*, *24*(2), 321–337.

Foerster, F., Borghi, A., & Goslin, J. (2020). Labels strengthen motor learning of new tools. *Cortex*, *129*, 1 – 10.
URL `http://www.sciencedirect.com/science/article/pii/S0010945220301477`

Foglia, L., & Wilson, R. A. (2013). Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(3), 319–325.

Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current directions in psychological science*, *24*(1), 17–23.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160.

Frith, C. (2003). What do imaging studies tell us about the neural basis of autism. *Autism: Neural basis and treatment possibilities*, (pp. 149–176).

Froudist-Walsh, S., Bliss, D. P., Ding, X., Jankovic-Rapan, L., Niu, M., Knoblauch, K., Zilles, K., Kennedy, H., Palomero-Gallagher, N., & Wang, X.-J. (2020). A dopamine gradient controls access to distributed working memory in monkey cortex. *bioRxiv*.

Fry, P. S. (1992). Assessment of private and inner speech of older adults in relation to depression. *Private speech: From social interaction to self-regulation*, (pp. 267–284).

Fuster, J. M., & Bressler, S. L. (2015). Past makes future: role of pfc in prediction. *Journal of cognitive neuroscience*, *27*(4), 639–654.

Galle, M. E., & McMurray, B. (2014). The development of voicing categories: A quantitative review of over 40 years of infant speech perception research. *Psychonomic bulletin & review*, *21*(4), 884–906.

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and cognition*, *17*(3), 887–910.

Gao, W.-J., & Mack, N. R. (2021). From hyposociability to hypersociability—the effects of psd-95 deficiency on the dysfunctional development of social behavior. *Frontiers in Behavioral Neuroscience*, *15*, 1.

Garagnani, M., & Pulvermüller, F. (2013). Neuronal correlates of decisions to speak and act: Spontaneous emergence and dynamic topographies in a computational model of frontal and temporal areas. *Brain and language*, *127*(1), 75–85.

Garagnani, M., Wennekers, T., & Pulvermüller, F. (2008). A neuroanatomically grounded hebbian-learning model of attention–language interactions in the human brain. *European Journal of Neuroscience*, *27*(2), 492–513.

Gazzaley, A., Clapp, W., Kelley, J., McEvoy, K., Knight, R. T., & D'Esposito, M. (2008). Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proceedings of the National Academy of Sciences*, *105*(35), 13122–13126.

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in cognitive sciences*, *16*(2), 129–135.

Geva, S., Jones, P. S., Crinion, J. T., Price, C. J., Baron, J.-C., & Warburton, E. A. (2011). The neural correlates of inner speech defined by voxel-based lesion–symptom mapping. *Brain*, *134*(10), 3071–3082.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A pdp model. *Cognitive psychology*, *44*(3), 297–337.

Gläscher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., Paul, L. K., & Tranel, D. (2012). Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, *109*(36), 14681–14686.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2008). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral cortex*, *19*(2), 483–495.

Gold, J. I., & Watanabe, T. (2010). Perceptual learning. *Current biology: CB*, *20*(2).

Goldman-Rakic, P. S. (1996). The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. *Philos Trans R Soc Lond B Biol Sci*, *351*(1346), 1445–1453.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science*, *1*(1), 69–78.

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep Learning*. Boston, MA: The MIT Press.
URL `http://www.iro.umontreal.ca/~bengioy/dlbook`

Gottlieb, J. (2007). From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron*, *53*(1), 9–16.
URL `http://dx.doi.org/10.1016/j.neuron.2006.12.009`

Grandin, T. (2009). How does visual thinking work in the mind of a person with autism? a personal account. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1522), 1437–1442.

Gregor, K., Papamakarios, G., Besse, F., Buesing, L., & Weber, T. (2019). Temporal difference variational auto-encoder. In *International Conference on Learning Representations*.
URL `https://openreview.net/forum?id=S1x4ghC9tQ`

Gruber, A. J., Dayan, P., Gutkin, B. S., & Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of computational neuroscience*, *20*(2), 153.

Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, *31*.

Hamby, D. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental monitoring and assessment*, *32*(2), 135–154.

Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching: Towards a unified developmental approach. *Developmental Science*, *13*(4), 622–635.

Happé, F., Booth, R., Charlton, R., & Hughes, C. (2006). Executive function deficits in autism spectrum disorders and attention-deficit/hyperactivity disorder: examining profiles across domains and ages. *Brain and cognition*, *61*(1), 25–39.

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.

Hartley, A. A., & Speer, N. K. (2000). Locating and fractionating working memory using functional neuroimaging: storage, maintenance, and executive functions. *Microscopy research and technique*, *51*(1), 45–53.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*, 245–258.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1601–1613.

Hearne, L. J., Mattingley, J. B., & Cocchi, L. (2016). Functional brain networks related to individual differences in human intelligence at rest. *Scientific reports*, *6*, 32328.

Heaton, R., Chelune, G., Talley, J., Kay, G., Curtiss, G., di Hardoy, M., Carta, M., Hardoy, M., & e Cabras, P. (2000). *WCST: Wisconsin card sorting test : forma completa revisionata : manuale.*. Firenze: O.S.

Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal anterior cingulate cortex: a bottom-up view. *Annual review of neuroscience*, *39*, 149–170.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012). A neurocomputational account of cognitive deficits in parkinson's disease. *Neuropsychologia*, *50*(9), 2290–2302.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, (pp. 599–619). Springer.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–1780.

Hoffmann, M. (2013). The human frontal lobes and frontal network systems: an evolutionary, clinical, and treatment perspective. *ISRN Neurol*, *2013*, 892459.

Holland, L., & Low, J. (2010). Do children with autism use inner speech and visuospatial resources for the service of executive control? evidence from suppression in dual tasks. *British journal of developmental psychology*, *28*(2), 369–391.

Holmes, A., Franklin, A., Clifford, A., & Davies, I. (2009). Neurophysiological evidence for categorical perception of color. *Brain and cognition*, *69*(2), 426–434.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554–2558.

Houk, J. C., Davis, J. L., & Beiser, D. G. (1995). *Models of information processing in the basal ganglia*. MIT press.

Huang, M.-H., Rust, R., & Maksimovic, V. (2019). The feeling economy: Managing in the next generation of artificial intelligence (ai). *California Management Review*, *61*(4), 43–65.

Humphreys, K., Hasson, U., Avidan, G., Minshew, N., & Behrmann, M. (2008). Cortical patterns of category-selective activation for faces, places and objects in adults with autism. *Autism Research*, *1*(1), 52–63.

Humphries, M. D., & Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in neurobiology*, *90*(4), 385–417.

Hurlburt, R. T., Alderson-Day, B., Kühn, S., & Fernyhough, C. (2016). Exploring the ecological validity of thinking on demand: neural correlates of elicited vs. spontaneously occurring inner speech. *PloS one*, *11*(2), e0147932.

Illing, B., Gerstner, W., & Brea, J. (2019). Biologically plausible deep learning—but how far can we go with shallow networks? *Neural Networks*, *118*, 90–101.

Impieri, D., Zilles, K., Niu, M., Rapan, L., Schubert, N., Galletti, C., & Palomero-Gallagher, N. (2019). Receptor density pattern confirms and enhances the anatomic-functional features of the macaque superior parietal lobule areas. *Brain Structure and Function*, *224*(8), 2733–2756.

Intaitė, M., Noreika, V., Šoliūnas, A., & Falter, C. M. (2013). Interaction of bottom-up and top-down processes in the perception of ambiguous figures. *Vision research*, *89*, 24–31.

Jaarsma, P., & Welin, S. (2012). Autism as a natural human variation: Reflections on the claims of the neurodiversity movement. *Health care analysis*, *20*(1), 20–30.

Jacob, S. N., & Nienborg, H. (2018). Monoaminergic neuromodulation of sensory processing. *Frontiers in neural circuits*, *12*, 51.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., & Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends Neurosci*, *18*(7), 314–320.

John-Steiner, V. (2014). Private speech among adults. In *Private speech*, (pp. 295–306). Psychology Press.

Johnson, K. A., Robertson, I. H., Kelly, S. P., Silk, T. J., Barry, E., Dáibhis, A., Watchorn, A., Keavey, M., Fitzgerald, M., Gallagher, L., et al. (2007). Dissociation in performance of children with adhd and high-functioning autism on a task of sustained attention. *Neuropsychologia*, *45*(10), 2234–2245.

Joseph, R. M., Steele, S. D., Meyer, E., & Tager-Flusberg, H. (2005). Self-ordered pointing in children with autism: failure to use verbal mediation in the service of working memory? *Neuropsychologia*, *43*(10), 1400–1411.

Jung, M., Matsumoto, T., & Tani, J. (2019). Goal-directed behavior under variational predictive coding: Dynamic organization of visual attention and working memory. Preprint arXiv 1903.04932v1.

Kaland, N., Smith, L., & Mortensen, E. L. (2008). Brief report: cognitive flexibility and focused attention in children and adolescents with asperger syndrome or high-functioning autism as measured on the computerized version of the wisconsin card sorting test. *Journal of autism and developmental disorders*, *38*(6), 1161–1165.

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*(4), 231–242.

Kaplan, G. B., Şengör, N. S., Gürvit, H., Genç, İ., & Güzeliş, C. (2006). A composite neural network model for perseveration and distractibility in the wisconsin card sorting test. *Neural Networks*, *19*(4), 375–387.

Kappel, D., Nessler, B., & Maass, W. (2014). Stdp installs in winner-take-all circuits an online approximation to hidden markov model learning. *PLoS Comput Biol*, *10*(3), e1003511.

Keehn, B., Müller, R.-A., & Townsend, J. (2013). Atypical attentional networks and the emergence of autism. *Neuroscience & Biobehavioral Reviews*, *37*(2), 164–183.

Kim, T., Hamade, K. C., Todorov, D., Barnett, W. H., Capps, R. A., Latash, E. M., Markin, S. N., Rybak, I. A., & Molkov, Y. I. (2017). Reward based motor adaptation mediated by basal ganglia. *Frontiers in computational neuroscience*, *11*, 19.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307–392. URL `http://dx.doi.org/10.1561/2200000056`

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.

Knoblauch, A., Körner, E., Körner, U., & Sommer, F. T. (2014). Structural synaptic plasticity has high memory capacity and can explain graded amnesia, catastrophic forgetting, and the spacing effect. *PloS one*, *9*(5), e96485.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307.

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, *11*(1), 16–22.

Koldewyn, K., Jiang, Y. V., Weigelt, S., & Kanwisher, N. (2013). Global/local processing in autism: Not a disability, but a disinclination. *Journal of autism and developmental disorders*, *43*(10), 2329–2340.

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature neuroscience*, *11*(2), 224–231.

Kornmeier, J., Hein, C. M., & Bach, M. (2009). Multistable perception: when bottom-up and top-down coincide. *Brain and cognition*, *69*(1), 138–147.

Kosslyn, S. M. (1999). *Image and brain*. Cambridge, MA: The MIT Press, fourth ed.

Kotz, S. A., Meyer, M., & Paulmann, S. (2006). Lateralization of emotional prosody in the brain: an overview and synopsis on the impact of study design. *Progress in brain research*, *156*, 285–294.

Koutník, J., Schmidhuber, J., & Gomez, F. (2014). Evolving deep unsupervised convolutional networks for vision-based reinforcement learning. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, (pp. 541–548).

Kray, J., Eber, J., & Lindenberger, U. (2004). Age differences in executive functioning across the lifespan: The role of verbalization in task preparation. *Acta Psychologica*, *115*(2-3), 143–165.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390–16395.

Kröger, B., Birkholz, P., Kannampuzha, J., & Neuschaefer-Rube, C. (2007). Modeling the perceptual magnet effect and categorical perception using self-organizing neural networks. In *Proceedings of the International Congress of Phonetic Sciences. Saarbrücken, Germany*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Brain and Behavioural Sciences*, *40*, 1–72.

Langland-Hassan, P., & Vicente, A. (2018). *Inner speech: New voices*. Oxford University Press, USA.

Laskin, M., Srinivas, A., & Abbeel, P. (2020). Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, (pp. 5639–5650). PMLR.

Le Roux, N., & Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, *20*(6), 1631–1649.

Levine, D. S., & Prueitt, P. S. (1989). Modeling some effects of frontal lobe damage—novelty and perseveration. *Neural networks*, *2*(2), 103–116.

Li, C.-S. R. (2004). Do schizophrenia patients make more perseverative than non-perseverative errors on the wisconsin card sorting test? a meta-analytic study. *Psychiatry Research*, *129*(2), 179–190.

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, *7*(6), 651–657.

Lie, C.-H., Specht, K., Marshall, J. C., & Fink, G. R. (2006). Using fmri to decompose the neural processes underlying the wisconsin card sorting test. *Neuroimage*, *30*(3), 1038–1049.

Lim, S.-J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in neuroscience*, *8*, 230.

Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine*, *4*(2), 4–22.

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703–713.

Lœvenbruck, H., Grandchamp, R., Rapin, L., Nalborczyk, L., & Dohen, M. (2018). A cognitive neuroscience view of inner language. *Inner speech: New voices*, (p. 131).

Loevenbruck, H., Grandchamp, R., Rapin, L., Perrone-Bertolotti, M., Pichat, C., Haldin, C., Cousin, E., Lachaux, J.-P., Dohen, M., Perrier, P., et al. (2019). Neural correlates of inner speaking, imitating and hearing: an fmri study. In *ICPhS 2019-19th International Congress of Phonetic Sciences*.

Lopez, B. R., Lincoln, A. J., Ozonoff, S., & Lai, Z. (2005). Examining the relationship between executive functions and restricted, repetitive symptoms of autistic disorder. *Journal of autism and developmental disorders*, *35*(4), 445–460.

Lupyan, G. (2005). Carving nature at its joints and carving joints into nature: How labels augment category representations. In *Modeling Language, Cognition And Action*, (pp. 87–96). World Scientific.

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, *24*(4), 279–284.

Lynn, C. W., & Bassett, D. S. (2019). The physics of brain network structure, function and control. *Nature Reviews Physics*, *1*(5), 318–332.

Maier, M., Glage, P., Hohlfeld, A., & Rahman, R. A. (2014). Does the semantic content of verbal categories influence categorical perception? an erp study. *Brain and Cognition*, *91*, 1–10.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, *32*(1), 4–18.

Mannella, F., & Baldassarre, G. (2015a). Selection of cortical dynamics for motor behaviour by the basal ganglia. *Biological Cybernetics*, *109*, 575–595.

Mannella, F., & Baldassarre, G. (2015b). Selection of cortical dynamics for motor behaviour by the basal ganglia. *Biological cybernetics*, *109*(6), 575–595.

Mannella, F., Gurney, K., & Baldassarre, G. (2013). The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. *Frontiers in Behavioral Neuroscience*, *7*(135), e1–29.

Mansouri, F. A., Matsumoto, K., & Tanaka, K. (2006). Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a wisconsin card sorting test analog. *Journal of Neuroscience*, *26*(10), 2745–2756.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. New York: Pantheon Books.

Martin, A. (2016). Grapes—grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic bulletin & review*, *23*(4), 979–990.

Marvel, C. L., & Desmond, J. E. (2012). From storage to manipulation: how the neural correlates of verbal working memory reflect varying demands on inner speech. *Brain and language*, *120*(1), 42–51.

McClelland, D. E., James L. andRumelhart, & the PDPResearchGroup (1986). *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1. Cambridge,MA: The MIT Press.

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216–271.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, vol. 24, (pp. 109–165). Elsevier.

McInroe, T. A., Spurrier, M., Sieber, J., & Conneely, S. (2021). Analyzing the hidden activations of deep policy networks: Why representation matters. *arXiv preprint arXiv:2103.06398*.

Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cereb Cortex*, *14*(11), 1256–1265.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature communications*, *11*(1), 1–12.

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science*, *372*(6545), 911–912.

Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends Neurosci*, *32*(7), 376–382.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167–202.

Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Prog Neurobiol*, *50*(4), 381–425.

Mirolli, M., Mannella, F., & Baldassarre, G. (2010). The roles of the amygdala in the affective regulation of body, brain and behaviour. *Connection Science*, *22*(3), 215–245.

Mirolli, M., & Parisi, D. (2006). Talking to oneself as a selective pressure for the emergence of language. In *The evolution of language*, (pp. 214–221). World Scientific.

Mollick, J. A., & Kober, H. (2020). Computational models of drug use and addiction: A review. *Journal of abnormal psychology*, *129*(6), 544.

Monchi, O., Petrides, M., Doyon, J., Postuma, R. B., Worsley, K., & Dagher, A. (2004). Neural bases of set-shifting deficits in parkinson's disease. *Journal of Neuroscience*, *24*(3), 702–710.

Monchi, O., Taylor, J. G., & Dagher, A. (2000). A neural model of working memory processes in normal subjects, parkinson's disease and schizophrenia for fmri design and predictions. *Neural Networks*, *13*(8-9), 953–973.

Mottron, L., Dawson, M., Soulieres, I., Hubert, B., & Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *Journal of autism and developmental disorders*, *36*(1), 27–43.

Murphy, C. M., Christakou, A., Daly, E. M., Ecker, C., Giampietro, V., Brammer, M., Smith, A. B., Johnston, P., Robertson, D. M., Consortium, M. A., et al. (2014). Abnormal functional activation and maturation of fronto-striato-temporal and cerebellar regions during sustained attention in autism spectrum disorder. *American Journal of Psychiatry*, *171*(10), 1107–1116.

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., Poldrack, R. A., Shohamy, D., & Kay, K. (2018). Cognitive computational neuroscience: A new conference for an emerging discipline. *Trends in cognitive sciences*, *22*, 365–367.

Neftci, E., Das, S., Pedroni, B., Kreutz-Delgado, K., & Cauwenberghs, G. (2014). Event-driven contrastive divergence for spiking neuromorphic systems. *Frontiers in neuroscience*, *7*, 272.

Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *12*(4), 313–324.

Niu, M., Impieri, D., Rapan, L., Funck, T., Palomero-Gallagher, N., & Zilles, K. (2020). Receptor-driven, multimodal mapping of cortical areas in the macaque monkey intraparietal sulcus. *Elife*, *9*, e55979.

Nolfi, S., & Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. Cambridge, MA: MIT Press.

Norman, D. (1988). *The Psychology of Everyday Things*. New York: Basic Books.

Nyhus, E., & Barceló, F. (2009). The wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain and cognition*, *71*(3), 437–451.

Ognibene, D., & Baldassarre, G. (2015). Ecological active vision: four bio-inspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Transactions on Autonomous Mental Development*, *7*(1), 3–25.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

O'Regan, J. K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav Brain Sci*, *24*(5), 939–73; discussion 973–1031.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283–328.

O'Regan, J. K., Myin, E., & Noë, A. (2005). Sensory consciousness explained (better) in terms of corporality and alerting capacity. *Phenomenology and the Cognitive Sciences*, *4*(4), 369–387.

O'Reilly, R., & Munakata, Y. (2000). Computational explorations in cognitive neuroscience-understanding the mind by simulating the brain. a bradford book.

Panksepp, J. (1998). *Affective neuroscience: the foundations of human and animal emotions*. Oxford: Oxford Unversity Press.

Paolo, A. M., Tröster, A. I., Axelrod, B. N., & Koller, W. C. (1995). Construct validity of the wcst in normal elderly and persons with parkinson's disease. *Archives of Clinical Neuropsychology*, *10*(5), 463–473.

Parisi, S., Ramstedt, S., & Peters, J. (2017). Goal-driven dimensionality reduction for reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (pp. 4634–4639). IEEE.

Parks, E. L., & Madden, D. J. (2013). Brain connectivity and visual attention. *Brain Connectivity*, *3*(4), 317–338.

Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: Metacognitive networks and measures of consciousness. *Cognition*, *117*(2), 182–190.

Passingham, R. E., & Wise, S. P. (2012). *The neurobiology of the prefrontal cortex: anatomy, evolution, and the origin of insight*, vol. 50. Oxford: Oxford University Press.

Pellicano, E. (2010). The development of core cognitive skills in autism: A 3-year prospective study. *Child Development*, *81*(5), 1400–1416.

Perani, D., Schnur, T., Tettamanti, M., Cappa, S. F., Fazio, F., et al. (1999). Word and picture matching: a pet study of semantic category effects. *Neuropsychologia*, *37*(3), 293–306.

Pérez-Gay, F., Christian, T., Gregory, M., Sabri, H., Harnad, S., & Rivas, D. (2017). How and why does category learning cause categorical perception? *International journal of comparative psychology*, *30*.

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M., & Loevenbruck, H. (2014). What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, *261*, 220–239.

Pessoa, L. (2015). Multiple influences of reward on perception and attention. *Visual cognition*, *23*(1-2), 272–290.

Peters, J. F., Tozzi, A., Ramanna, S., & İnan, E. (2017). The human brain from above: an increase in complexity from environmental stimuli to abstractions. *Cognitive neurodynamics*, *11*(4), 391–394.

Peters, R. (2006). Ageing and the brain. *Postgraduate medical journal*, *82*(964), 84–88.

Petrolini, V., Jorba, M., & Vicente, A. (2020). The role of inner speech in executive functioning tasks: Schizophrenia with auditory verbal hallucinations and autistic spectrum conditions as case studies. *Frontiers in Psychology*, *11*, 2452.

Pleger, B., Blankenburg, F., Ruff, C. C., Driver, J., & Dolan, R. J. (2008). Reward facilitates tactile judgments and modulates hemodynamic responses in human primary somatosensory cortex. *Journal of Neuroscience*, *28*(33), 8161–8168.

Pleger, B., Ruff, C. C., Blankenburg, F., Klöppel, S., Driver, J., & Dolan, R. J. (2009). Influence of dopaminergically mediated reward on somatosensory decision-making. *PLoS biology*, *7*(7), e1000164.

Poort, J., Khan, A. G., Pachitariu, M., Nemri, A., Orsolic, I., Krupic, J., Bauza, M., Sahani, M., Keller, G. B., Mrsic-Flogel, T. D., et al. (2015). Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron*, *86*(6), 1478–1490.

Posner, I. (2020). Robots thinking fast and slow: On dual process theory and metacognition in embodied ai. *Robotics Retrospectives Workshop 2020*.

Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in human neuroscience*, *9*, 197.

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not 'grandmother-cell'coding in the medial temporal lobe. *Trends in cognitive sciences*, *12*(3), 87–91.

Raffone, A., Srinivasan, N., & van Leeuwen, C. (2014). The interplay of attention and consciousness in visual search, attentional blink and working memory consolidation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130215.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*(4), 1009–1023.

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131.

Rentzeperis, I., Nikolaev, A. R., Kiper, D. C., & van Leeuwen, C. (2014). Distributed processing of color and form in the visual cortex. *Frontiers in psychology*, *5*, 932.

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370–379.

Rigotti, M., Ben Dayan Rubin, D. D., Wang, X.-J., & Fusi, S. (2010). Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Frontiers in computational neuroscience*, *4*, 24.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu Rev Neurosci*, *27*, 169–192.

Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: anatomy and functions. *Exp Brain Res*, *153*(2), 146–157.

Robertson, C. E., & Baron-Cohen, S. (2017). Sensory perception in autism. *Nature Reviews Neuroscience*, *18*(11), 671–684.

Robertson, C. E., Thomas, C., Kravitz, D. J., Wallace, G. L., Baron-Cohen, S., Martin, A., & Baker, C. I. (2014). Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain*, *137*(9), 2588–2599.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in cognitive sciences*, *16*(3), 147–156.

Rumsey, J. M. (1985). Conceptual problem-solving in highly verbal, nonretarded autistic men. *Journal of autism and developmental disorders*, *15*(1), 23–36.

Russell-Smith, S. N., Comerford, B. J., Maybery, M. T., & Whitehouse, A. J. (2014). Brief report: Further evidence for a link between inner speech limitations and executive function in high-functioning children with autism spectrum disorders. *Journal of autism and developmental disorders*, *44*(5), 1236–1243.

Salminen, N. H., Tiitinen, H., & May, P. J. (2009). Modeling the categorical perception of speech sounds: A step toward biological plausibility. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(3), 304–313.

Samanez-Larkin, G. R., & Knutson, B. (2015). Decision making in the ageing brain: changes in affective and motivational circuits. *Nature Reviews Neuroscience*, *16*(5), 278–289.

Sanada, M., Ikeda, K., Kimura, K., & Hasegawa, T. (2013). Motivation enhances visual working memory capacity through the modulation of central cognitive processes. *Psychophysiology*, *50*(9), 864–871.

Santucci, V. G., Baldassarre, G., & Mirolli, M. (2016). Grail: A goal-discovering robotic architecture for intrinsically-motivated learning. *8*(3), 214–231.

Schendan, H. E., & Ganis, G. (2012). Electrophysiological potentials reveal cortical mechanisms for mental imagery, mental simulation, and grounded (embodied) cognition. *Frontiers in psychology*, *3*, 329.

Schmidt, R. A. (1976). The schema as a solution to some persistent problems in motor learning theory. In *Motor control*, (pp. 41–65). Elsevier.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*(2), 461–464.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews*, *32*(2), 265–278.

Seger, C. A., & Miller, E. K. (2010a). Category learning in the brain. *Annual review of neuroscience*, *33*, 203–219.

Seger, C. A., & Miller, E. K. (2010b). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219.

Shao, L., Zhu, F., & Li, X. (2014). Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, *26*(5), 1019–1034.

Shu, B.-C., Lung, F.-W., Tien, A. Y., & Chen, B.-C. (2001). Executive function deficits in non-retarded autistic children. *Autism*, *5*(2), 165–174.

Sidtis, J. J., Van Lancker Sidtis, D., Dhawan, V., & Eidelberg, D. (2018). Switching language modes: complementary brain patterns for formulaic and propositional language. *Brain connectivity*, *8*(3), 189–196.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(6869), 318–320.

Silvetti, M., Vassena, E., Abrahamse, E., & Verguts, T. (2018). Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLoS computational biology*, *14*, e1006370.

Sinzig, J., Bruning, N., Morsch, D., & Lehmkuhl, G. (2008). Attention profiles in autistic children with and without comorbid hyperactivity and attention problems. *Acta Neuropsychiatrica*, *20*(4), 207–215.

Siu, C. R., & Murphy, K. M. (2018). The development of human visual cortex and clinical implications. *Eye and brain*, *10*, 25.

Sivagnanam, S., Majumdar, A., Yoshimoto, K., Astakhov, V., Bandrowski, A. E., Martone, M. E., & Carnevale, N. T. (2013). Introducing the neuroscience gateway. *IWSG*, *993*.

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, *28*.

Soulières, I., Mottron, L., Saumier, D., & Larochelle, S. (2007). Atypical categorical perception in autism: Autonomy of discrimination? *Journal of autism and developmental disorders*, *37*(3), 481–490.

Sperati, V., & Baldassarre, G. (2018). A bio-inspired model learning visual goals and attention skills through contingencies and intrinsic motivations. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 326–344.

Spratling, M. W., & Johnson, M. H. (2006). A feedback model of perceptual learning and categorization. *Visual Cognition*, *13*(2), 129–165.

Squire, L., Berg, D., Bloom, F. E., Du Lac, S., Ghosh, A., & Spitzer, N. C. (Eds.) (2012). *Fundamental neuroscience*. Academic Press.

Srivastava, R. K., Masci, J., Kazerounian, S., Gomez, F., & Schmidhuber, J. (2013). Compete to compute. In *Advances in neural information processing systems*, (pp. 2310–2318).

Steinke, A., Lange, F., & Kopp, B. (2020a). Parallel model-based and model-free reinforcement learning for card sorting performance. *Scientific reports*, *10*(1), 1–18.

Steinke, A., Lange, F., Seer, C., Hendel, M. K., & Kopp, B. (2020b). Computational modeling for neuropsychological assessment of bradyphrenia in parkinson's disease. *Journal of Clinical Medicine*, *9*(4), 1158.

Steinke, A., Lange, F., Seer, C., & Kopp, B. (2018). Toward a computational cognitive neuropsychology of wisconsin card sorts: a showcase study in parkinson's disease. *Computational Brain & Behavior*, *1*(2), 137–150.

Stewart, T., & Eliasmith, C. (2011). Neural cognitive modelling: A biologically constrained spiking neuron model of the tower of hanoi task. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33.

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience*, *29*(5), 1565–1572.

Stuss, D., Levine, B., Alexander, M., Hong, J., Palumbo, C., Hamer, L., Murphy, K., & Izukawa, D. (2000). Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, *38*(4), 388–402.

Sullivan, E. V., Adalsteinsson, E., Hedehus, M., Ju, C., Moseley, M., Lim, K. O., & Pfefferbaum, A. (2001). Equivalent disruption of regional white matter microstructure in ageing healthy men and women. *Neuroreport*, *12*(1), 99–104.

Sun, J., Wang, X., Xiong, N., & Shao, J. (2018). Learning sparse representation with variational auto-encoder for anomaly detection. *IEEE Access*, *6*, 33353–33361.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: The MIT Press, second edition, in progress ed.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*. MIT press.

Taffoni, F., Focaroli, V., Keller, F., & Iverson, J. M. (2019). Motor performance in a shape sorter task: A longitudinal study from 14 to 36 months of age in children with an older sibling asd. *PloS one*, *14*(5), e0217416.

Tajima, C. I., Tajima, S., Koida, K., Komatsu, H., Aihara, K., & Suzuki, H. (2016). Population code dynamics in categorical perception. *Scientific reports*, *6*(1), 1–13.

Thill, S., Caligiore, D., Borghi, A. M., Ziemke, T., & Baldassarre, G. (2013). Theories and computational models of affordance and mirror systems: An integrative review. *Neuroscience and Biobehavioral Reviews*, *37*, 491–521.

Thomas, V., Bengio, E., Fedus, W., Pondard, J., Beaudoin, P., Larochelle, H., Pineau, J., Precup, D., & Bengio, Y. (2018). Disentangling the independently controllable factors of variation by interacting with the world. *arXiv preprint arXiv:1802.09484*.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and brain sciences*, *16*(3), 495–511.

Tommasino, P., Caligiore, D., Mirolli, M., & Baldassarre, G. (2019). A reinforcement learning architecture that transfers knowledge between skills when solving multiple tasks. *IEEE Transactions on Cognitive and Developmental Systems*, *11*(2), 292–317.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol Bull*, *215*(3), 216–242.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci*, *17*(7), 450–461.

Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science*, *363*(6428), 692–693.

Van Geit, W., De Schutter, E., & Achard, P. (2008). Automated neuron model optimization techniques: a review. *Biological cybernetics*, *99*, 241–251.

Vernon, D. (2008). Cognitive vision: The case for embodied perception. *Image and Vision Computing*, *26*(1), 127–140.

Vickery, T. J., Chun, M. M., & Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron*, *72*(1), 166–177.

Volpato, C., Schiff, S., Facchini, S., Silvoni, S., Cavinato, M., Piccione, F., Antonini, A., & Birbaumer, N. (2016). Dopaminergic medication modulates learning from feedback and error-related negativity in parkinson's disease: a pilot study. *Frontiers in behavioral neuroscience*, *10*, 205.

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, *20*(2), 150–159.

Wakita, M. (2004). Categorical perception of orientation in monkeys. *Behavioural processes*, *67*(2), 263–272.

Wallace, G. L., Silvers, J. A., Martin, A., & Kenworthy, L. E. (2009). Brief report: Further evidence for inner speech deficits in autism spectrum disorders. *Journal of autism and developmental disorders*, *39*(12), 1735.

Wang, D., & Gu, J. (2018). Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, *16*(5), 320–331.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860–868.

Wang, T., Lavis, Y., Hall, G., & Mitchell, C. J. (2012). Location and salience of unique features in human perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(4), 407.

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychol Rev*, *66*, 297–333.

Whorf, B. L. (2012). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. Mit Press.

Wilcox, T., Woods, R., & Chapa, C. (2008). Color–function categories that prime infants to use color information in an object individuation task. *Cognitive Psychology*, *57*(3), 220–261.

Williams, D. M., Bowler, D. M., & Jarrold, C. (2012). Inner speech is used to mediate short-term memory, but not planning, among intellectually high-functioning adults with autism spectrum disorder. *Development and psychopathology.*, *24*(1), 225–239.

Williams, D. M., Peng, C., & Wallace, G. L. (2016). Verbal thinking and inner speech use in autism spectrum disorder. *Neuropsychology Review*, *26*(4), 394–419.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3-4), 229–256.

Williams, S. M., & Goldman-Rakic, P. S. (1993). Characterization of the dopaminergic innervation of the primate frontal cortex using a dopamine-specific antibody. *Cerebral Cortex*, *3*(3), 199–222.

Wills, A. J., Suret, M., & McLaren, I. P. (2004). The role of category structure in determining the effects of stimulus preexposure on categorization accuracy.

Winsler, A., Abar, B., Feder, M. A., Schunn, C. D., & Rubio, D. A. (2007). Private speech and executive functioning among high-functioning children with autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*(9), 1617–1635.

Witzel, C., & Gegenfurtner, K. R. (2016). Categorical perception for red and brown. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(4), 540.

Woldorff, M. G., Fox, P., Matzke, M., Lancaster, J., Veeraswamy, S., Zamarripa, F., Seabolt, M., Glass, T., Gao, J., Martin, C., et al. (1997). Retinotopic organization of early visual spatial attention effects as revealed by pet and erps. *Human brain mapping*, *5*(4), 280–286.

Wolters, G., & Raffone, A. (2008). Coherence and recurrency: Maintenance, control and integration in working memory. *Cognitive Processing*, *9*(1), 1–17.

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The contrast database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour*, *6*(4), 593–604.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321.

Yin, H. H., & Knowlton, B. J. (2006a). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476.

Yin, H. H., & Knowlton, B. J. (2006b). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476.

Yin, H. H., Ostlund, S. B., & Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience*, *28*(8), 1437–1448.

Zacks, J. M. (2008). Neuroimaging studies of mental rotation: a meta-analysis and review. *Journal of cognitive neuroscience*, *20*(1), 1–19.

Zald, D. H., & Andreotti, C. (2010). Neuropsychological assessment of the orbital and ventromedial prefrontal cortex. *Neuropsychologia*, *48*(12), 3377–3391.

Zanolie, K., Teng, S., Donohue, S. E., van Duijvenvoorde, A. C., Band, G. P., Rombouts, S. A., & Crone, E. A. (2008). Switching between colors and shapes on the basis of positive and negative feedback: An fmri and eeg study on feedback-based learning. *cortex*, *44*(5), 537–547.

Zanto, T. P., Rubens, M. T., Thangavel, A., & Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature neuroscience*, *14*(5), 656–661.

Zappacosta, S., Mannella, F., Mirolli, M., & Baldassarre, G. (2018). General differential hebbian learning: Capturing temporal relations between events in neural networks and the brain. *Plos Computational Biology*, *14*(8), e1006227.

Zarr, N., & Brown, J. (2019). Computational neural mechanisms of goal-directed planning and problem solving. *bioRxiv*, (p. 779306).

Zenke, F., Agnes, E. J., & Gerstner, W. (2015). Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature communications*, *6*(1), 1–13.

Zhang, Q.-s., & Zhu, S.-C. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 27–39.